Analyse de la diversité microbienne par séquençage massif : méthodes et applications

Najwa Taïb

Numéro d'ordre : D.U. 2374

UNIVERSITÉ BLAISE PASCAL

U.F.R Sciences et Technlogies

École Doctorale des Sciences de la Vie, Santé, Agronomie et Environnement

Numéro 614

THÈSE

Présentée pour obtenir le grade de

DOCTEUR D'UNIVERSITÉ

Spécialité : Génétique, Physiologie et Bioinformatique

par **Najwa Taïb**

Analyse de la diversité microbienne par séquençage massif : Méthodes et Applications

Soutenue publiquement le 29/8/2013, devant le jury composé de :

Rapporteurs	Eric Coissac	MCF, Université Joseph Fourier Grenoble
	Jean-François Humbert	DR. INRA, ENS, Paris (Président)
Membres et invités	Thomas Pommier	CR. INRA, Lyon
	Gisèle Bronner	MCF, Université Clermont Ferrand II
	Didier Debroas	Prof, Université Clermont Ferrand II
	Engelbert Mephu-Nguifo	Prof, Université Clermont Ferrand II

Remerciements

A l'issue de ces quatre ans, je suis convaincue que la thèse est loin d'être un travail solitaire. En effet, je n'aurais jamais pu réaliser ce travail doctoral sans le soutien d'un grand nombre de personnes dont la générosité, la bonne humeur et l'intérêt manifestés à l'égard de ma recherche m'ont permis de progresser dans cette phase délicate de «l'apprenti-chercheur».

En premier lieu, je tenais à remercier les deux personnes sans qui ce travail n'aurait pu voir le jour sans leurs encadrements, leurs conseils et leurs soutiens tout au long de l'élaboration de cette thèse. Merci à mes chefs, Didier et Gisèle, pour votre confiance en acceptant d'encadrer ce travail, et pour avoir su être là lorsque les difficultés pointaient du nez.

Je souhaiterai également exprimer ma gratitude à Christian Amblard et Télesphore Sime-Ngando, les deux directeurs respectifs du LMGE pour m'avoir accueillie au sein de leur laboratoire.

Eric Coissac et Jean-François Humbert m'ont fait l'honneur d'être rapporteurs de ma thèse, ils ont pris le temps de m'écouter et de discuter avec moi. Leurs remarques m'ont permis d'envisager mon travail sous un autre angle. Pour tout cela, je les remercie. Merci également aux autres membres du jury : Thomas Pommier, Engelbert Mephu-Nguifo et Vincent Breton, pour l'intérêt qu'ils ont porté à mon travail.

Au cours de ces années, j'ai fait partie de l'équipe « Microbiologie de l'environnement et Bioinformatique ». Les discussions que j'ai pu avoir durant les réunions d'équipe ou en dehors avec les différents membres m'ont beaucoup apportée. Je remercie donc Mylène, car ensemble nous avons appris que le bout du tunnel est au fond ... ; ainsi que Jean-Christophe, Agnès, Corinne, Isa J, Anne, Viviane, Isa M, Emilie, Simon et François pour leurs conseils, mais également pour les moments de détente durant les pauses gourmandes ..

En dehors des MEB, je tiens à remercier Aurélie et Diane pour nos discussions super profondes et hautement philosophiques ..

Un grand merci à Amine et Khaled, pour leurs conseils, leurs encouragements et leur patience quand ça n'allait pas .. mais également pour nos escapades furtives loin de la ville ..

Enfin, je ne pourrai finir sans exprimer ma grande gratitude à mes Parents, sans vous rien de tout cela n'aurait eu lieu .. littéralement ... Merci ! !

Résumé

Les avancées des nouvelles techniques de séquençage (NGS) ont permis dans le cadre des études en écologie microbienne de passer de l'analyse de quelques centaines de séquences par étude à des centaines de millions de séquences. Cette différence quantitative des données produites a induit des différences qualitatives quant aux études réalisées. En effet, avec le changement du type de données, les approches classiques d'analyse ne peuvent être appliquées et il est devenu nécessaire de définir de nouvelles stratégies en tenant compte des contraintes que posent ces données. Alors qu'il était possible d'insérer classiquement quelques dizaines de séquences issues des techniques de première génération dans des phylogénies expertisées, le nombre de séquences généré aujourd'hui par les NGS à chaque expérience rend cette tâche irréalisable et nécessite la mise en place de nouvelles stratégies et l'utilisation d'outils adaptés. Par ailleurs, les outils disponibles d'analyse de la diversité microbienne adaptés aux amplicons de nouvelle génération, implémentent des approches probabilistes et/ou de recherche de similitude pour l'identification des séquences environnementales. L'approche phylogénétique quant à elle, bien qu'elle soit la plus robuste, n'est pas utilisée pour l'annotation taxonomique de ce type de données du fait de ses besoins en temps et en ressources de calcul. Au-delà de l'approche d'annotation taxonomique, les nouvelles techniques de séquençage posent également le problème de la qualité des séquences produites et son impact sur l'estimation de la diversité. Ainsi, ce travail de thèse avait pour objectif la définition d'une stratégie d'analyse bioinformatique de données de séquençage massif dans le contexte de l'étude de la diversité microbienne, en tenant compte des limitations imposées par les ressources informatiques actuelles (matérielles et logicielles) d'un côté, et de l'avantage des méthodes phylogénétiques par rapport aux autres approches d'annotation taxonomique. Ce travail a donné lieu au développement d'une chaîne de traitement proposant une série d'analyses allant des séquences brutes jusqu'à la visualisation des résultats, tout en replaçant les séquences environnementales dans un contexte évolutif. L'approche développée a été optimisée pour la gestion de gros volumes de données, et a été comparée en terme de précision d'affiliation aux autres approches communément utilisées en écologie microbienne. Les tests et simulations ont montré qu'à partir d'une taille d'amplicons de 400 pb, l'affiliation phylogénétique avait

les meilleurs résultats mais aussi, que la qualité de cette affiliation différait selon la région hypervariable ciblée. La chaîne de traitements mise en place a ensuite été par implémentée dans un contexte de calcul à haute performance, notamment sur un cluster de calcul, pour proposer un service web dédié à l'analyse de la diversité microbienne.

Abstract

The characterization of microbial community structure via SSU rRNA gene profiling has been greatly advanced in recent years by the introduction of NGS amplicons, leading to a better representation of sample diversity at a lower cost. This progress in method development has provided a new window into the composition of microbial communities and sparked interest in the members of the rare biosphere. Concurrently, the processing of such amount of data has become an important bottleneck for the effectiveness of microbial ecology studies, and a multitude of analysis platforms have been developed for the handling of these data. As implemented, these tools have a steep learning curve for the biologist who is not computationally inclined, as they require extensive user intervention and consume many CPU hours due to dataset analysis and complexity, which can present a significant barrier to researchers. Moreover, although phylogenetic affiliation has been shown to be more accurate for the taxonomic assignment of NGS reads, the existing tools assign taxonomy by either a similarity search or a probabilistic approach, with the phylogenies being restricted to samples' comparison. Beyond the taxonomic assignment, the new sequencing technologies also arise the problem of the quality of the generated sequences and its impact on the richness estimation. In this work, we aimed to define a strategy for the bioinformatic analysis of high-throughput sequences in order to depict the microbial diversity, taking into account both the limitations imposed by current computer resources (hardware and software), and the advantage of the phylogenetic methods over the other taxonomic annotation approaches. This work has led to the development of a pipeline offering a set of analyzes ranging from raw sequences processing to the visualization of the results, while replacing the environmental sequences in an evolutionary framework. The developed approach was optimized for managing large volumes of data, and has been compared in term of the accuracy of taxonomic assignment to the approaches commonly used in the field of microbial ecology. This pipeline was then used to the developement of a dedicated web server for high-throughput sequencing analysis, that relies on a computing cluster and performs large-scale phylogeny-based analyses of rRNA genes with no need for specialized informatics expertise, and uses the phylogenies for both the taxonomy assessment and the delineation of monophyletic groups to highlight clades of interest.

Table des matières

1 Synthèse bibliographique **17**

1.1 Introduction 19

1.2 Apports du séquençage massif à l'écologie microbienne 21

1.2.1 Métagénomique et Métagénétique 21

1.2.2 La notion d'espèce : de la microbiologie à la microbiologie de l'environnement 25

1.2.3 La structure des communautés microbiennes 26

1.3 Méthodes liées au traitement des données de la métagénétique 28

1.3.1 Contraintes de la définition d'une OTU en métagénétique 29

1.3.2 Le regroupement des séquences en OTUs : richesse et abondance . . 30

1.3.3 L'annotation taxonomique : analyse de la composition 36

1.4 Limites de la métagénétique en écologie microbienne 42

1.4.1 Impact des erreurs de séquençage sur la richesse 42

1.4.2 Impact de la profondeur de séquençage sur les mesures de diversité 46

1.5 Objectifs de l'étude et organisation du mémoire 47

2 Méthodologie **51**

2.1 Introduction 53

2.2 Choix méthodologiques 55

2.2.1 Alignement 55

2.2.2 Phylogénie 56

2.2.3 Recherche de similitude 57

2.3 Calcul distribué et parallélisation 59

2.3.1 Grille de calcul 59

2.3.2 Cluster de calcul 61

Article 1 **65**

Article 2 **87**

3 Applications **93**

3.1 Introduction 95

3.2 Erreurs de séquençage : qualité et nettoyage 95

3.3 Normalisation 98

3.4 Seuils de similitude et OTUs rares 99

3.5 Etude des clades 101

Article 3 **102**

Article 4 **121**

4 Discussion **145**

4.1 PANAM et l'affiliation taxonomique 148

4.1.1 Impact de la région et de la taille de l'amplicon sur les assignations 148

4.1.2 Identification des clades 150

4.1.3 Impact des bases de référence sur les assignations 151

4.1.4 Le problème de la notion d'espèce en microbiologie 152

4.2 PANAM et les biais de l'estimation de la richesse et de la diversité 153

4.2.1 Les biais dans l'estimation des indices de richesse et de diversité . . 153

4.2.2 Les biais dans l'estimation de la richesse 155

4.3 Conclusion et Perspectives 158

Bibliographie **161**

Annexe **182**

Table des figures

1.1 Représentation graphique du développement du matériel informatique de stockage par rapport aux outils de séquençage. 20
1.2 Représentation schématique de la variabilité le long de l'ADNr. 30
1.3 Schématisation des trois stratégies de la clusterisation hiérarchique. 32
1.4 Schématisation de l'approche de la clusterisation gourmande implémentée dans UCLUST. 33

2.1 Représentation graphique de la spécificité de l'affiliation phylogénétique en fonction de la taille des séquences et du rang taxonomique. 54
2.2 Histogramme représentant la spécificité moyenne de l'affiliation de BLAST et UCLUST calculée sur 4 x 1000 séquences complètes. 58
2.3 Représentation graphique du découpage des tâches de PANAM en 4 services WPE. 60
2.4 Captures d'écran de ePANAM, correspondant au formulaire de saisie (a), les graphes générés pour l'alpha-diversité (b) et la béta-diversité (c). 62

3.1 Types et fréquences des mutations par variant par rapport à la séquence de référence. 97
3.2 Les valeurs des scores de qualité en fonction du type de mutation (a) et du taux d'erreur (b). 98
3.3 La phylogénie des 1948 séquences de *B. hominis* avec des séquences de référence. 101

4.1 Le nombre et la nature des mutations par séquence sur deux jeux de données non nettoyés générés à partir de la séquence de *B. hominis* par deux plates-formes différentes. 157

Liste des tableaux

1.1 Comparaison des caractéristiques des différentes plates-formes de séquençage massif. 24
1.2 Comparaison des caractéristiques des principaux outils de clusterisation utilisés en écologie microbienne. 35
1.3 Caractéristiques des outils utilisés en écologie microbienne pour l'annotation taxonomique des séquences d'ARNr 16S et 18S. 40

2.1 Comparaison des temps de traitement de différents outils d'alignement pour l'insertion d'une séquence dans deux profils de taille différente. 56
2.2 Comparaison des topologies des phylogénies générées par différents outils par rapport au maximum de vraisemblance. 57
2.3 Temps moyens d'exécution des différents services de PANAM sur la grille pour différents jeux de séquences 61

3.1 Nombre et structure des OTUs générées par différentes méthodes à différents seuils de similitude pour les 1948 séquences nettoyées de *B. hominis*. 100

Chapitre 1

Synthèse bibliographique

1.1 Introduction

Le rôle des micro-organismes dans le fonctionnement général de la biosphère est connu depuis longtemps (Falkowski et al., 2008), mais de façon surprenante, leur diversité spécifique et fonctionnelle, les mécanismes régissant leur dispersion et leur histoire évolutive demeurent encore mal compris. Au cours des deux dernières décennies, l'essor spectaculaire des approches moléculaires (e.g., PCR, séquençage Sanger, empreintes génétiques (Fuhrman and Hagström, 2008)) a permis d'approfondir ces questions, et d'accéder à une fraction de micro-organismes restée jusqu'alors inaccessibles par les techniques culturales. Cependant, bien que les outils moléculaires mettaient en évidence l'existence d'une grande diversité taxonomique (Pace, 1997) et métabolique (Venter et al., 2004) parmi les espèces les plus abondantes, ils ne permettaient toujours pas de détecter une majorité d'espèces dites "rares" (Pedrós-Alió, 2006).

L'apparition et la généralisation des techniques de séquençage de nouvelle génération (NGS) ont révolutionné les études en écologie microbienne, en autorisant l'accès à cette fraction microbienne rare et la réalisation de projets de recherche de grande envergure (e.g., le consortium du microbiome humain (Turnbaugh et al., 2007), ou encore l'expédition GOS et l'échantillonnage des océans à l'échelle du globe (Venter et al., 2004)). En effet, alors que le séquençage Sanger dit de première génération produit quelques centaines de séquences d'une longueur d'environ 1000 pb, les NGS génèrent 10^6-10^9 séquences d'une taille allant de 100 à 700 pb à un coût plus faible (Glenn, 2011, Scholz et al., 2012). Néanmoins, malgré l'évolution permanente de ces NGS, leur limite réside encore aujourd'hui dans la taille des fragments séquencés qui est relativement courte, ce qui restreint l'information biologique qu'ils portent. De plus, l'émergence des NGS étant récente, les biais pouvant découler de leur utilisation font toujours l'objet de plusieurs études (Quince et al., 2009, Reeder and Knight, 2010).

Avec l'accélération de l'usage des technologies NGS dès le milieu des années 2000, les chercheurs ont dû faire face à des quantités colossales de données, nécessitant pour les stocker et les traiter un changement d'échelle des ressources informatiques (Figure 1.1). De nouveaux outils d'analyse bio-informatique adaptés à ce type de données ont également dû être développés. A titre d'exemple, la technique Illumina HiSeq2500 peut produire jusqu'à 600 Gigaoctets (Go) de données ce qui correspond à $6x10^9$ séquences

(Scholz et al., 2012) et nécessite environ 0.6 Téraoctets (To) d'espace sur un disque de stockage (Glenn, 2011). Actuellement, pour analyser ce type de données (e.g., annotation, assemblage), les ressources informatiques minimales requises consistent en un processeur à 4 cœurs, 16 Gigaoctets de RAM et 2 Téraoctets d'espace disque (Logares et al., 2012).

Pour des analyses répétées, des supports informatiques adaptés sont également nécessaires, comme des structures de calcul distribuées (e.g., clusters (ou grappes), grilles de calcul, cloud [1]).

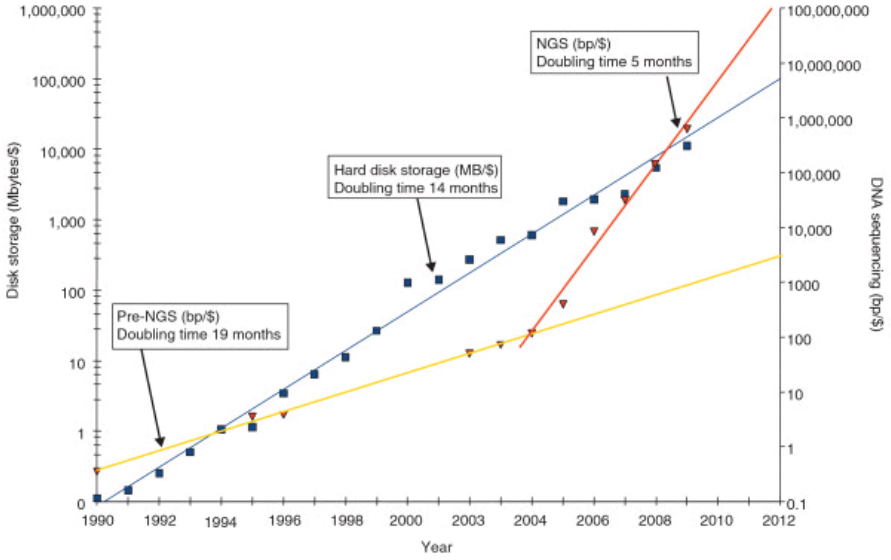


Figure 1.1. Représentation graphique du développement du matériel informatique de stockage par rapport aux outils de séquençage.

Les coûts du matériel informatique de stockage (en bleu) sont exprimés en dollars par paire de bases, par rapport aux techniques de première génération (en jaune) et aux techniques de nouvelle génération (en rouge).

Outre les besoins croissants en ressources informatiques dont le développement a été dépassé par le développement des techniques de séquençage (Stein et al., 2010), la gestion de ces grandes quantités de données pose également le problème d'outils d'analyse, avec la nécessité de développements méthodologiques. Actuellement, la disponibilité d'outils génériques d'analyse de la diversité microbienne, faciles d'utilisation, accessibles et bien

1. Le cloud computing désigne l'utilisation d'un ensemble de serveurs distants pour traiter et/ou stocker l'information., accessible via un navigateur web.

documentés sous forme de pipelines [2], est devenue cruciale pour les scientifiques. En écologie microbienne, les données de séquençage sont traitées soit par des packages [3] d'analyse publiques mettant à disposition plusieurs outils de traitement (e.g., QIIME (Caporaso et al., 2010)), soit par des workflow [4] ré-implémentant ces outils avec la possibilité de les chaîner (e.g., Mothur (Schloss et al., 2009)). Toutefois, la limite de cette stratégie réside dans les besoins croissants en temps de développement et en support technique, ainsi que dans les erreurs résultant des ré-implémentations de chaque étape. Une autre stratégie consiste à proposer une plate-forme permettant le chaînage de différents outils pour répondre à un besoin particulier et de mettre ces chaînes de traitement à la disposition de la communauté (e.g., Galaxy (Goecks et al., 2010)). D'autres outils, tels que Pyrotagger (Kunin and Hugenholtz, 2010), QIIME EC2, CLoVR-16S (Angiuoli et al., 2011) se basent sur des machines virtuelles ou des clouds pour effectuer des calculs intensifs.

Alors que la révolution actuelle en écologie microbienne semble se diriger vers une synergie entre les techniques de séquençage massif et les ressources informatiques matérielles et logicielles, il devient nécessaire de définir des chaînes de traitement qui prennent en compte d'une part les limitations imposées par le volume des données, et d'autre part les biais introduits par les séquenceurs à haut débit.

1.2 Apports du séquençage massif à l'écologie microbienne

1.2.1 Métagénomique et Métagénétique

Durant la dernière décennie, les technologies de séquençage de l'ADN ont connu une profonde transformation et une démocratisation (Shendure and Ji, 2008). Alors que seuls les laboratoires ayant accès à des ressources substantielles pouvaient auparavant s'engager dans des projets de séquençage massif pour décrire des communautés microbiennes complexes, aujourd'hui la réduction des coûts de séquençage autorise un plus grand nombre de

2. un procédé de traitement qui contient différentes étapes indépendantes les unes des autres
3. un ensemble de programmes réalisés pour une application précise
4. une suite de tâches chaînées

laboratoires à produire massivement des séquences avec les outils communément appelés « séquenceurs de paillasse » (e.g., le junior de Roche) capables de générer jusqu'à 60 Mb nucléotides par heure (Loman et al., 2012). La possibilité d'accéder à une telle profondeur de séquençage pour des coûts réduits par les différentes plates-formes de séquençage (Tableau 1.1), a profondément bouleversé l'envergure des projets de séquençage proposés et la façon d'aborder les problématiques d'écologie microbienne.
Cette révolution a permis d'initier des séquençages aléatoires des génomes de l'ensemble des micro-organismes présents dans un environnement donné. Grâce à cette approche appelée métagénomique, il est désormais possible d'avoir une vision plus intégrative de l'ensemble des évènements se déroulant dans un écosystème (accès aux gènes codant les ARNr et les ARN fonctionnels et aux gènes protéiques). L'article fondateur de cette méthode est celui de Venter et al. (2004). Depuis, cette stratégie a été appliquée pour l'étude de différents écosystèmes tels que le sol (Fierer et al., 2007), les microbiomes intestinaux (Qin et al., 2012) ou encore les lacs (Debroas et al., 2009).

Plus récemment, les technologies NGS ont été utilisées pour le séquençage d'amplicons (produits d'amplification de la PCR (Polymerase Chain Reaction)) correspondant aussi bien à des marqueurs phylogénétiques que fonctionnels. Cette stratégie nommée métagénétique par Bik et al., a émergé avec les travaux pionniers menés par Sogin et al. (2006) ciblant le gène codant l'ARN ribosomique (ADNr 16S pour les procaryotes et 18S pour les eucaryotes) obtenu à partir d'échantillons de milieux marins. Cette étude a démontré le potentiel de la métagénétique à haut débit dans l'évaluation de la richesse des micro-organismes par rapport aux outils moléculaires de première génération. Dans les études de métagénétique, chaque séquence obtenue pour le gène considéré représente un individu et sert à son identification. Dans ce type d'étude, les marqueurs phylogénétiques utilisés doivent permettre de retracer les liens de parenté entre les micro-organismes et avoir un modèle d'évolution qui reflète l'évolution des espèces dont ils proviennent. En l'occurrence, le gène codant pour la petite sous-unité de l'ARNr, transmis de manière verticale et ayant un rôle structural et catalytique dans le ribosome (Lawrence, 1999, Amann et al., 1995), est particulièrement efficace pour la reconstruction phylogénétique (Pace, 1997) ; il est de ce fait le marqueur phylogénétique le plus utilisé actuellement en écologie microbienne (Konopka, 2006). Bien que la métagénétique soit principalement appliquée au séquençage de l'ADNr 16S et 18S provenant de différents milieux (e.g., sols (Roesch

et al., 2007) ; milieux marins (Stoeck et al., 2010) ; lacs (Medinger et al., 2010) ; intestins (Turnbaugh et al., 2008a,b)), quelques études rapportent son application pour d'autres marqueurs (e.g., trnL (Valentini et al., 2009) ; la cytochrome oxydase ou COI (Hajibabaei et al., 2011)).

En terme d'inventaire de la diversité, la métagénomique et la métagénétique peuvent avoir des résultats qui divergent, et ce notamment à cause du transfert horizontal des gènes fonctionnels. Cependant, des études récentes ont montré que ces deux approches peuvent être complémentaires et avoir des résultats convergents (Turnbaugh et al., 2008b, Fierer et al., 2012, Harris et al., 2012). Par ailleurs, bien que la métagénétique décrive uniquement la composition taxonomique d'un environnement, elle permet également d'identifier des populations distinctes qui peuvent être associées à des écotypes et présenter des caractéristiques adaptées à des habitats différents (Ward et al., 2007).

Parmi les deux approches moléculaires présentées ci-dessus nous nous focaliserons par la suite sur la métagénétique en décrivant et discutant les nouveaux apports de cette méthode à la connaissance de la structure des communautés microbiennes et les méthodologies permettant de traiter ce type d'expérimentation.

Table 1.1. Comparaison des caractéristiques des différentes plates-formes de séquençage massif.

Plates-formes	Temps de production	Longueur des lectures (pb)	Nombre max de séquences	Volume de données	Type d'erreurs	Taux d'erreur(%)
Roche						
454 GS FLX+	23 heures	600-800	$1x10^6$	$<$ 700 Mb	Indel	1
454 GS FLX Titanium	10 h	400-500	$1\ 10^6$	$<$ 500 Mb	Indel	1
454 GS Junior	10 h	400-450	$1x10^5$	$\approx$35 Mb	Indel	1
Illumina						
HiSeq 2000	11 jours	100-200	$6x10^9$	$<$ 540-600 Gb	Substitution	$>$ 0.1
HiSeq 1000	8.5 j	100-200	$3x10^9$	$<$ 270-300 Gb	Substitution	$>$ 0.1
GAIIx	7.5-14.5 j	50-75	$6.4x10^8$	$<$ 95 Gb	Substitution	$>$ 0.1
MiSeq	19-24h	100-150	$7x10^6$	$<$ 1-2 Gb	Substitution	$>$ 0.1
Life technologies						
AB SOLiD 5500 system	4 j	35-75	$2.4x10^9$	$\approx$100 Gb	Biais A-T	$>$ 0.06
AB SOLiD 5500 xl system	7-8 j	35-75	$6x10^9$	$\approx$250 Gb	Biais A-T	$>$ 0.01
Ion torrent						
PGM 314-chip	3.5 h	100-200	$1x10^6$	$>$ 10 Mb	Indel	1
PGM 316-chip	4.7 h	100-200	$6x10^6$	$>$ 100 Mb	Indel	1
PGM 318-chip	5.5h	100-200	$11x10^6$	$>$ 1 Gb	Indel	1
Pacific biosciences						
RS	0.5 h	$>$ 1500	$50x10^3$	$\approx$60-75 Mb	Insertion	15

D'après Shokralla et al. (2012) et Scholz et al. (2012)

1.2.2 La notion d'espèce : de la microbiologie à la microbiologie de l'environnement

Le concept le plus largement utilisé pour définir une espèce est celui de l'espèce biologique proposée par Mayr (1942), bien que dans la réalité de nombreuses espèces (eucaryotes) soient définies sur la base de différences phénotypiques, les croisements étant difficilement observables dans la nature. L'application du concept biologique de l'espèce trouve ses limites dans la caractérisation des espèces microbiennes du fait de l'absence de reproduction sexuée associée à l'existence de transferts horizontaux entre individus évolutivement distants. Aussi, pour les procaryotes, les espèces sont actuellement définies de façon pragmatique en prenant en compte à la fois des caractères génotypiques et phénotypiques (Achtman and Wagner, 2008). Lorsque les traits phénotypiques ne peuvent être décrits, ce qui est le cas pour la majorité des micro-organismes sur terre, une désignation provisoire de l'espèce candidate peut être proposée sur la base d'une discrimination exclusivement génétique (*Candidatus sp.*). Celle-ci est généralement basée sur l'analyse de la similitude entre des séquences de la petite sous unité de l'ARN ribosomique. Le pourcentage d'identité de ce marqueur phylogénétique est ainsi utilisé pour regrouper les séquences dans des unités taxonomiques opérationnelles ou OTUs (Operational Taxonomic Unit). La majorité des études de diversité basées sur l'ADNr 16S définissent les OTUs comme étant des ensembles de séquences présentant au moins 97% d'identité entre elles. Ce seuil a été défini par Stackebrandt and Goebel (1994) comme équivalent au seuil d'hybridation ADN-ADN de 70% observé dans des expériences de réassociation réalisée entre les membres d'espèces bactériennes préétablies, issues d'organismes mis en culture.

La notion d'espèce est souvent vue comme l'unité de base de la biodiversité. Bien que ne définissant pas *in fine* des espèces, les OTUs sont devenues l'unité de mesure de la richesse spécifique en écologie microbienne.

1.2.3 La structure des communautés microbiennes

La détermination de la structure des communautés via la caractérisation de la richesse [5], l'abondance [6], la diversité et la composition spécifique [7] d'un écosystème est un enjeu central en écologie et donc en écologie microbienne. Dans les écosystèmes aquatiques les estimations en terme de richesse étaient jusqu'à récemment de moins de 200 OTUs bactériennes pour 90% des banques de clones [8] (Kemp and Aller, 2004). Par une approche de métagénétique à haut débit menée en milieux marins, Sogin et al. (2006), en détectant entre 1184 et 3290 OTUs, ont mis en évidence la sous-estimation de cette richesse spécifique. Par ailleurs, la plus grande partie de la biodiversité est représentée par des OTUs faiblement abondantes, c'est à dire qui ne peuvent être que rarement détectées par les techniques moléculaires de première génération (empreintes génétiques, séquençage Sanger). Ces espèces qualifiées de rares forment une distribution en longue queue dans les courbes de rang-abondance. Présentes en grand nombre dans la plupart des écosystèmes, elles constitueraient jusqu'à 75% de la biomasse. Patterson (2009) propose deux modèles pour expliquer la distribution des espèces rares au sein des écosystèmes aquatiques : i) chaque habitat comprend un nombre important de niches microbiennes renfermant une communauté très complexe, diversifiée et active ii) seule une faible proportion des espèces rares présentes est métaboliquement active, le reste se composant de formes en dormance parvenues à l'écosystème par dispersion. Pour cet auteur, le deuxième modèle prévaudrait dans le monde microbien pour lequel les grandes capacités de dispersion façonneraient la diversité présente. Cependant, le premier concept ne peut être exclu, Campbell et al. (2011) ayant récemment mis en évidence que ces espèces rares pouvaient être actives. Quoiqu'il en soit, il est admis que, dans l'environnement, la croissance bactérienne varie en fonction des changements de conditions environnementales (nutriments, température,....) ; et cette dépendance pourrait être à l'origine de remaniements saisonniers entre espèces rares et dominantes. Ainsi, il est vraisemblable que les fluctuations des micro-organismes qui mettent en évidence des populations dominantes à certaines saisons qui ne sont plus

5. le nombre d'espèces différentes
6. le nombre d'individus pour une espèce donnée
7. l'association de la richesse et de l'abondance des espèces dans un environnement
8. ensemble de fragments d'ADN

détectées par la suite traduisent des remaniements entre la fraction dominante et rare qui n'ont pu être mis en évidence par les méthodes classiques (Fuhrman et al., 2006, Boucher et al., 2006) mais qui auraient pu l'être par les méthodes à haut débit.

Selon une étude sur la dynamique saisonnière des communautés bactériennes en milieu marin réalisée par Caporaso et al. (2011b), la grande majorité des taxa détectée à une saison donnée serait représentée par au moins une séquence durant les saisons suivantes si la profondeur de séquençage était accrue (des millions de lectures au lieu de milliers). Contrairement aux idées reçues, les variations de la composition des communautés bactériennes dans le temps et dans l'espace observées jusqu'à maintenant (Sogin et al., 2006, Fuhrman, 2009, Kirchman et al., 2010, Caporaso et al., 2011a), ne seraient donc pas dues à la présence ou à l'absence de groupes taxonomiques, mais plutôt à une fluctuation de l'abondance relative de certains taxa de la biosphère rare et de groupes dominants (Campbell et al., 2011). Caron and Countway (2009) suggèrent que cette hypothèse serait également applicable aux communautés de protistes. Cependant, la nature des facteurs environnementaux à l'origine des modifications substantielles entre dominants et rares restent à déterminer. En revanche, ce modèle d'alternance rares/dominants s'oppose à une autre vision découlant des travaux de Galand et al. (2009) sur les *Archaea*, qui montrent que certains de ces micro-organismes détectés comme étant rares, restent toujours rares quelles que soient les conditions environnementales et que, de plus, ils sont spécifiques d'une aire géographique. L'hypothèse de ces auteurs est que les taxa rares seraient actifs à faible abondance.

Bien que l'ampleur de la diversité des communautés microbiennes demeure un sujet de débat actuel, une distribution de fréquence à longue queue de la communauté microbienne semble être cohérente et réelle (Morales et al., 2009). En effet la biosphère rare, qui comporte également des organismes en dormance et pouvant se réactiver quand les conditions sont favorables (seed bank ou banque de graines) (Jones and Lennon, 2010), aurait des implications sur notre compréhension des limites de la résistance de l'environnement aux perturbations (Groffman et al., 2006). Ce concept est notamment décrit dans les écosystèmes marins par O'Dor et al. (2010), où les micro-organismes peuvent diffuser dans tout l'océan pendant des millions d'années, créant ainsi un réservoir d'espèces permettant de

maintenir un équilibre face aux changements environnementaux et de stabiliser l'écosystème.

La réalité et l'importance de la biosphère rare font l'objet de plusieurs débats, de nombreux travaux suggérant une surestimation de celle-ci due à des artefacts méthodologiques (Reeder and Knight, 2010, Quince et al., 2009, Huse et al., 2010). On entend généralement par OTUs rares, des OTUs dont l'abondance est inférieure à 1-0.1 et même 0.01% de la diversité totale détectée dans un environnement, cette valeur différant en fonction des études. La détermination de cette biodiversité est fortement dépendante de la définition d'une OTU, des erreurs dans les séquences (PCR et séquençage), de la profondeur de séquençage et de la méthode d'affiliation. **Ainsi, la suite du manuscrit aura pour objectifs de présenter les différentes méthodes utilisées pour la description de la diversité microbienne à partir des données de métagénétique à haut débit, et les limites de cette approche.**

1.3 Méthodes liées au traitement des données de la métagénétique

Le problème majeur dans les études de communautés en écologie microbienne réside dans l'identification des micro-organismes présents dans l'échantillon. Cependant, toutes les étapes employées pour obtenir les données, à savoir l'extraction de l'ADN, l'amplification d'un gène cible, le regroupement des séquences dans des OTUs et l'assignation d'une taxonomie à chaque OTU sont sujettes à des erreurs et des biais méthodologiques (Hamady and Knight, 2009), ce qui peut générer de nombreuses erreurs et donc fausser la description des communautés microbiennes. De plus, le changement d'échelle du volume de données à traiter a induit le changement des méthodes de calcul et des ressources informatiques classiques par des outils plus adaptés (Gonzalez and Knight, 2012). En effet, de nombreux outils et algorithmes pour le traitement des données de métagénétique ont été développés ; et leurs résultats peuvent conduire à des interprétations différentes quant à la richesse et l'abondance des organismes présents dans un environnement donné (Liu et al., 2008, Kunin et al., 2010, Quince et al., 2009).

1.3.1 Contraintes de la définition d'une OTU en métagénétique

Comme nous l'avons vu, les études de diversité basées sur les gènes marqueurs analysent souvent la composition des communautés microbiennes en terme d'OTUs, et ce pour pallier à l'ambiguïté de la définition d'une espèce en microbiologie (Cohan, 2002). Le seuil traditionnellement utilisé pour regrouper les séquences d'une même espèce en OTUs est 97%. Or, ce seuil ne peut pas être appliqué à tous les organismes et ce pour différentes raisons. D'une part parce qu'il a été défini à partir d'organismes isolés en culture et ceux-ci ne représentent qu'une minorité des micro-organismes présents dans l'environnement, d'autre part parce qu'il a été défini sur la séquence complète de la petite sous unité de l'ADN ribosomique 16S alors que la grande majoritée des séquences générées, notamment avec les NGS sont de taille inférieure. Or, l'information biologique portée dépend de la taille du fragment et de la région amplifiée. En effet, la présence de régions hypervariables (v1-v9) (Goebel and Stackebrandt, 1994) le long de la petite sous unité de l'ARN ribosomique, chacune avec des taux de variabilité différents (Figure 1.2), conduit à des résolutions taxonomiques différentes (Liu et al., 2007, 2008, Wang et al., 2007). Il est donc nécessaire d'adapter les seuils de clusterisation en OTUs à la région amplifiée. D'après une étude réalisée par Kim et al. (2011), le seuil de clusterisation varie de 96% à 98% pour l'ARNr 16S selon la région étudiée ; et ils préconisent l'utilisation des régions v1-v3 et v1-v4 pour une meilleure affiliation des bactéries, et de la région v1-v3 pour les *Archaea*. Pour les eucaryotes, Caron et al. (2009) ont étudié les variations inter- et intra-spécifiques sur le gène codant l'ARNr 18S, et ont conclu à un seuil de 95% pour clusteriser les séquences de protistes au niveau de l'espèce. Chez ces derniers, ce sont les régions v4 et v9 qui sont les plus utilisées pour des études de diversité (Behnke et al., 2011, Stoeck et al., 2010, Pawlowski et al., 2011).

Une méthode alternative pour la définition de la notion d'espèce, est l'intégration des informations phylogénétiques et des relations évolutives entre les organismes, et ce dans le but de s'affranchir des biais dûs à la vitesse d'évolution différentielle entre les espèces. En effet, selon Koeppel and Wu (2013), les séquences contenues dans une OTU générée selon un seuil d'identité ne sont pas forcément monophylétiques, ce qui peut se traduire par une hétérogénéité écologique dans une OTU. Ces auteurs préconisent par ailleurs de remplacer les OTUs par la notion d'écotypes, qui au-delà de définir une unité pour les mesures de la diversité microbienne, incorporent les modèles d'évolution et peuvent être

discriminés sur la base de paramètres environnementaux.

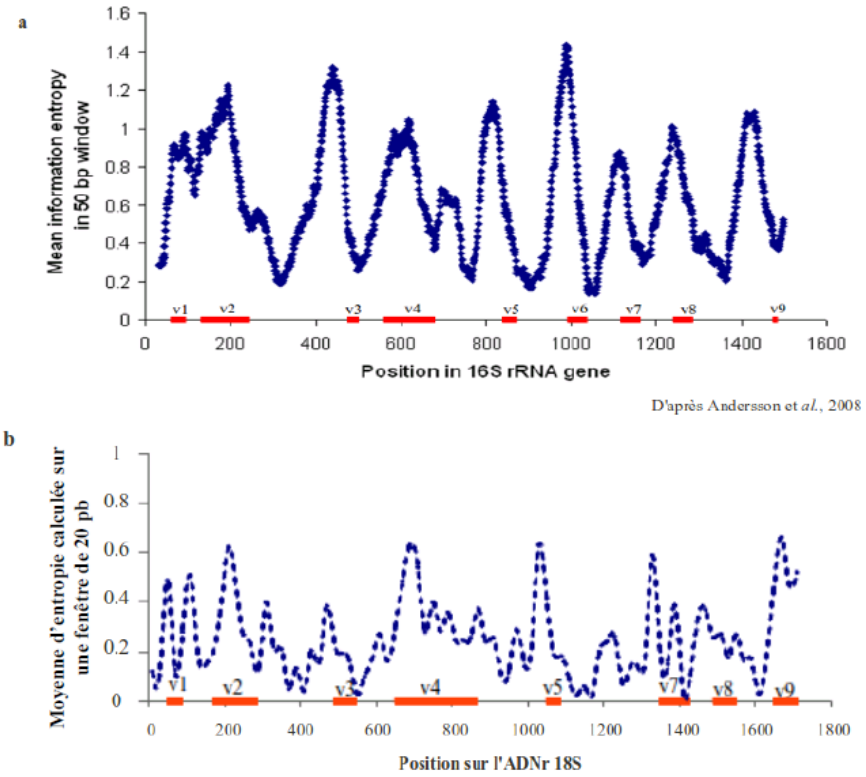


Figure 1.2. Représentation schématique de la variabilité le long de l'ADNr. La variabilité est exprimée par l'entropie de Shannon ($H(x) = -\sum_{i=1}^{n} P_i log(P_i)$) en fonction des régions hypervariables (en rouge) sur les gènes codant pour l'ARNr 16S (a) (Andersson et al., 2008) et 18S (b).

1.3.2 Le regroupement des séquences en OTUs : richesse et abondance

Le regroupement des séquences en OTUs par une approche de classification non supervisée permet d'énumérer d'éventuels nouveaux organismes non cultivés. En effet, plutôt que de regrouper tous les organismes qui n'ont pas d'équivalents proches connus dans les bases de référence sous l'étiquette « non classés » ou « inconnus », la clusterisation permet de comparer les séquences les unes par rapport aux autres pour ensuite les regrouper dans des clusters incluant une certaine part de variabilité (définie par le seuil de clusterisation). Ainsi, le nombre et la taille des OTUs sont utilisés pour déterminer la diversité de l'échantillon étudié indépendamment de toute identification taxonomique.

Cependant, outre le choix du seuil de clusterisation pour définir les OTUs, le choix de la méthode de regroupement des séquences dans des clusters peut avoir un impact sur

le nombre et la taille des OTUs générées. En effet, le choix d'un seuil n'indique pas la manière de recruter des séquences dans un cluster, d'autant plus que pour avoir des résultats semblables en terme de richesse ou d'abondance, le seuil n'est pas le même suivant que la séquence est ajoutée à un cluster quand elle est similaire à toutes les séquences s'y trouvant ou seulement à la séquence consensus du cluster (Schloss and Handelsman, 2005, Schloss and Westcott, 2011).

De nombreux algorithmes de clusterisation existent, et ils peuvent être classés en trois groupes (Cheng et al., 2012) :

- la clusterisation hiérarchique (HC pour Hierachical Clustering) est implémentée dans de nombreux outils tels que Mothur (Schloss et al., 2009) ou encore ESPRIT (Sun et al., 2009). La figure 1.3 illustre les trois stratégies de la clusterisation hiérarchique, à savoir la clusterisation par liaison simple (SL pour Single Linkage) connue aussi sous le nom de la méthode du plus proche voisin (NN pour Nearest Neighbor), par liaison complète (CL pour Complete Linkage) connue aussi par la méthode du voisin le plus éloigné (FN pour Furthest Neighbor) et la classification ascendante hiérarchique par la moyenne arithmétique (AN pour Average Neighbor). Dans les études d'écologie microbienne employant les techniques de séquençage de première génération, la méthode FN était utilisée, car elle garantit que les séquences regroupées au sein d'un cluster soient distantes au maximum du seuil défini ; alors que pour les deux autres approches, les séquences regroupées dans les OTUs peuvent partager des identités inférieures au seuil fixé. Ces approches basées sur le calcul des distances entre toutes les séquences sont coûteuses en temps de calcul et en taille mémoire en particulier pour construire la matrice de distance. Pour ce qui est du calcul des distances, celles-ci peuvent correspondre à des distances d'édition [9] ou des distances basées sur les k-mer [10].

9. Distance d'édition : la distance d(a,b) entre deux séquences a et b correspond au nombre minimum d'éditions (mutation, insertion, délétion) nécessaires pour transformer a en b.
10. Distance basée sur les K-mers : le nombre des mots de k lettres commun entre deux séquences.

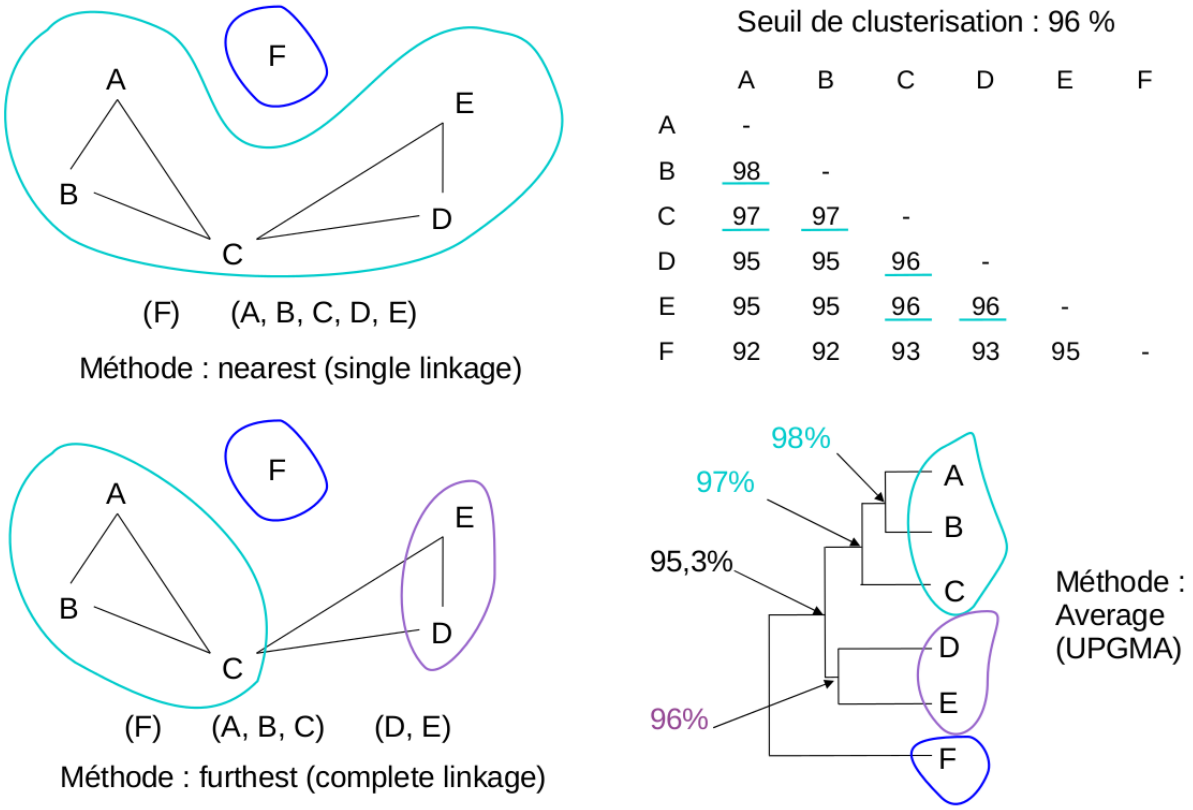
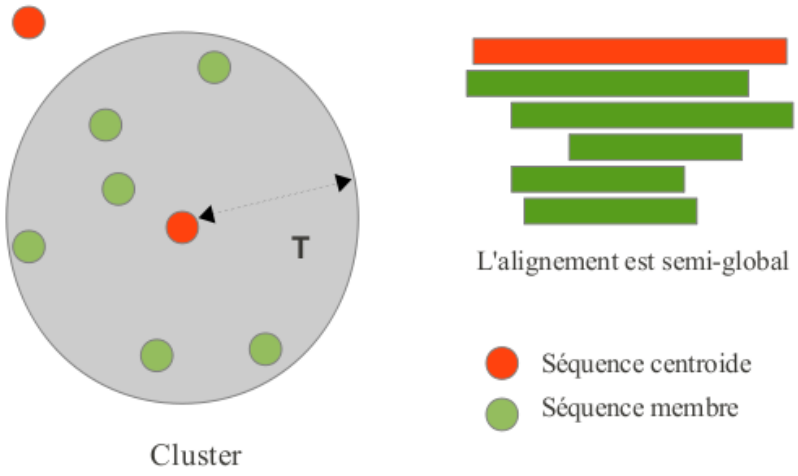


Figure 1.3. Schématisation des trois stratégies de la clusterisation hiérarchique.
Pour un seuil de 96%, les lignes entre les lettres indiquent l'existence d'une relation entre les séquences à ce seuil. Les lettres sont regroupées en fonction de la méthode utilisée. Pour la méthode FN, il existe deux groupements, ABC et CDE. Cependant, ABCDE n'ont pas été groupées car les distances entre (A,B) et (D,E) sont supérieures au seuil défini.

– la clusterisation hiérarchique gourmande (GHC pour Greedy Hierarchical Clustering) implémentée dans UCLUST (Edgar, 2010) et CD-HIT (Li et al., 2001). Il s'agit d'une clusterisation hiérarchique heuristique, c'est-à-dire qu'elle fournit rapidement une solution réalisable, mais pas nécessairement optimale. Comme les résultats sont dépendants de l'ordre dans lequel les séquences sont évaluées, ces dernières sont d'abord triées par taille et la première séquence (seq1) est considérée comme le centroïde du premier cluster (cluster1). Ensuite, le reste des séquences est comparé de manière séquentielle à chaque cluster, si la distance entre une séquence et un cluster est inférieure au seuil, la séquence est incorporée au cluster, sinon, un nouveau cluster est créé (figure 1.4). Cet algorithme a été développé afin de proposer des méthodes pouvant supporter de gros volumes de données et réduire la complexité

des calculs [11] et les besoins en ressources informatiques.



D'après http ://www.drive5.com

Figure 1.4. Schématisation de l'approche de la clusterisation gourmande implémentée dans UCLUST.
Le premier cluster testé dont le centroïde partage avec la séquence requête un pourcentage d'identité inférieur ou égale au seuil fixé, intégrera cette séquence.

- la clusterisation Bayesienne (BC pour Bayesian Clustering) implémentée dans CROP pour les séquences de l'ADNr 16S (Hao et al., 2011). Cette approche ne se base pas sur des pourcentages de similitude des séquences pour la clusterisation, mais sur une classification bayésienne pour regrouper les séquences. Cette classification utilise une liste de mots d'une longueur donnée (k-mers) présents dans une base d'apprentissage pour construire un modèle statistique, et hiérarchise les séquences requêtes en fonction de la probabilité de présence d'un k-mer pris au hasard sur cette séquence. Ces classificateurs produisent généralement moins d'OTUs que les approches HC et GHC.

Bien que largement utilisés par les microbiologistes (e.g., (Campbell et al., 2011, Harris

11. En informatique, la complexité décrit la « difficulté » à résoudre un problème. Elle est exprimée en fonction de la taille des données.

et al., 2012)), certains de ces outils n'ont été testés que sur des jeux de données limités et avec des paramètres spécifiques. Par ailleurs, une étude évaluant l'impact des choix méthodologiques sur le regroupement des séquences dans des clusters a été réalisée par White et al. (2010), et a démontré que l'utilisation en amont de paramètres différents, telle que la méthode d'alignement, par un même algorithme entraîne des résultats différents. Enfin, de la même manière que les régions hypervariables, le seuil utilisé pour regrouper les séquences dans des clusters dépend également de l'algorithme employé. Le choix du même seuil quels que soient la méthode et l'algorithme de clusterisation peut résulter en des estimations erronées des OTUs. Sun et al. (2012) préconisent par ailleurs la validation de ces choix par l'ajout de séquences connues dans le jeu de données qui permettraient de donner une approximation plus précise du seuil de clusterisation à utiliser. Le tableau 1.2 résume les caractéristiques de certains outils de clusterisation utilisés en écologie microbienne.

Table 1.2. Comparaison des caractéristiques des principaux outils de clusterisation utilisés en écologie microbienne.

Outil	Fonctions	Méthode d'alignement	Méthode de Clusterisation	Utilisation de bases données	Matrice de distance	Complexité du calcul [12]
DOTUR	Clusterisation	N/A	HC	Non	Oui	$O(N^2)$
Mothur	Alignement de séquences, Clusterisation	Alignement de profil, Alignement multiple, Alignement par paires de séquences	HC	Oui	Oui	$O(N^2)$
ESPRIT	Alignement de séquences, Clusterisation	Alignement par paires de séquences	HC	Non	Oui	$O(N^2)$
ESPRIT-Tree	Alignement de séquences, Clusterisation	Alignement par paires de séquences	HC	Non	Non	$O(N^{1.2})$
RDP/pyro	Alignement de séquences, Clusterisation	Infernal	HC	Oui	Oui	$O(N^2)$
CD-HIT	Alignement de séquences Clusterisation	Alignement par pair de séquences	GHC	Non	Non	$O(N^{1.2})$
UCLUST	Alignement de séquences, Clusterisation	Alignement par paires de séquences	GHC	Non	Non	$O(N^{1.2})$
CROP	Clusterisation,	Alignement par paires de séquences	BC	Non	Non	$O(N^2/k)$

Adapté de Sun et al. (2012)

[12]Compléxité du calcul : Un problème portant sur N données et appartenant à la classe de compléxité $O(N^k)$ sera résolu en un temps polynomial de paramètre k. Ainsi si N=100 et k=3, il faudra 10^6 opérations pour résoudre un tel problème

1.3.3 L'annotation taxonomique : analyse de la composition

Méthodes d'annotation

Alors que la classification *de novo* des séquences indépendamment de la taxonomie permet de regrouper les organismes dans des OTUs, les placer dans un contexte taxonomique permet de les relier à ce qui est connu dans les bases de référence. Dans le cadre des données issues des séquenceurs de nouvelle génération, deux approches sont souvent utilisées pour l'assignation taxonomique. Le BLAST (Altschul et al., 1997), se basant sur une recherche de similitude par le calcul des scores d'alignements locaux par paires de séquences, est le plus répandu dans le contexte des NGS et le plus simple d'utilisation. En effet, une seule étape permet d'affecter aux séquences expérimentales les séquences qui leur sont le plus proches parmi toutes celles présentes dans les bases. Le BLAST permet de comparer une séquence expérimentale contre les bases publiques (e.g., GenBank ; EMBL ; SILVA (Pruesse et al., 2007)), ou des bases construites par l'utilisateur et contenant les séquences d'organismes d'intérêt. Toutefois, cette approche basée sur la recherche de similitude présente plusieurs inconvénients : quand la séquence requête ne possède pas de séquences proches dans la base de référence, BLAST donne dans tous les cas un alignement, même contre des séquences relativement éloignées (c'est particulièrement le cas pour la petite sous unité de l'ADNr qui est conservée entre espèces) ce qui peut fausser l'affiliation. De plus, les valeurs utilisées par BLAST pour déterminer la similitude entre les séquences (e-value et score de similitude) peuvent être difficiles à interpréter car dépendant de la longueur de l'alignement (cas des scores) et des propriétés de la base de référence (cas de la e-value). Plusieurs outils dédiés à l'analyse des données du séquençage massif intègrent le BLAST dans leur chaîne de traitement (e.g. QIIME (Caporaso et al., 2010) ; CANGS (Pandey et al., 2010) ; Mothur (Schloss et al., 2009) ; WATERS (Hartman et al., 2010)).

Les approches probabilistes, tel que le classificateur RDP, implémentent une méthode de classification bayésienne. Cette méthode découpe les séquences expérimentales en « mots » ou k-mers de huit caractères, et utilise une base de référence comme base d'apprentissage pour trouver les « mots » correspondants aux séquences expérimentales (Cole et al., 2009, 2011). Elle se base sur le théorème de Bayes permettant de calculer les probabilités conditionnelles, et cherche la classification qui maximise la probabilité d'observer

les mots ou k-mers de la séquence requête. Lors de la phase d'apprentissage, le classificateur calcule les probabilités qu'une nouvelle séquence appartienne à telle catégorie à partir de la proportion des séquences d'apprentissage appartenant à cette catégorie. A chaque affiliation est associé un score reflétant la fiabilité de la taxonomie assignée. Là aussi, l'apprentissage peut se faire avec une base publique ou une base personnelle. Par ailleurs, la richesse et la taille de la base d'apprentissage, couplées à une affiliation taxonomique précise des séquences de référence améliorent la qualité des affiliations bayésiennes ; certains auteurs préconisent de limiter les séquences de la base à la région amplifiée pour une meilleure classification (Werner et al., 2011).

QIIME et Mothur intègrent également le classificateur RDP. De par leur facilité d'utilisation, le BLAST et le « classificateur RDP » sont largement utilisés dans les études de diversité en écologie microbienne pour affilier des taxonomies aux données du séquençage massif (Campbell et al., 2011, Cheung et al., 2010, Vila-Costa et al., 2013). Toutefois, les bases d'apprentissage implémentées par défaut dans ces outils sont retreintes aux séquences de l'ADNr 16S. Par conséquent l'annotation des séquences de l'ADNr 18S nécessitent la construction de bases de données appropriées.

Les méthodes basées sur les phylogénies constituent une approche alternative pour l'affiliation taxonomique et sont nécessaires pour l'identification de nouveaux taxa (Liu et al., 2008). En effet, cette approche a l'avantage de prendre en compte les relations de parenté entre différents organismes. Alors que le BLAST ou le classificateur RDP comparent une séquence expérimentale aux séquences de référence en termes d'identité ou de distance k-mers en la plaçant dans un large groupe taxonomique, la phylogénie permet d'analyser la séquence dans un contexte évolutif ; le branchement de la séquence requête indiquant ses liens par rapport à chaque séquence dans l'arbre. De plus, elle permet d'identifier les séquences qui ne possèdent pas de séquences proches dans les bases de référence, et de décrire d'éventuels nouveaux clades. Bien que coûteuse en terme de temps de calcul, la phylogénie reste une approche de choix pour l'analyse évolutive des séquences ; elle possède de solides bases statistiques d'inférence, des tests pour l'estimation de l'incertitude et des modèles évolutifs avancés (Matsen et al., 2010). Toutefois, cette approche est peu adaptée aux données du séquençage massif, notamment parce que les reconstructions phylogénétiques sont difficiles à automatiser et les phylogénies basées sur

le maximum de vraisemblance sont un problème NP-complet [13], les arbres ne pouvant être générés dans un laps de temps raisonnable par les ressources informatiques actuellement disponibles. Récemment, des outils ont été développés dans le but de proposer des algorithmes adaptés à de grands jeux de données (e.g., FastTree (Price et al., 2009, 2010) ; RAxML (Stamatakis, 2006, Stamatakis et al., 2008)). De plus, bien qu'il y ait eu beaucoup de progrès pour la représentation des arbres comportant des milliers de séquences (e.g., Dendroscope ; iToL (Letunic and Bork, 2007, 2011)), la comparaison et l'interprétation de telles phylogénies restent compliquées. Le tableau 1.3 résume les approches d'annotation taxonomique implémentées dans quelques outils de traitement de séquences utilisés en écologie microbienne.

Afin de contourner les problèmes liés à l'application des méthodes phylogénétiques classiques à de grands jeux de données, une approche alternative, le placement phylogénétique a été développée. En se basant sur une phylogénie de référence dans laquelle les séquences expérimentales sont insérées sans réévaluation de la topologie de l'arbre, le placement phylogénétique réduit considérablement le temps de traitement, d'autant plus que l'introduction des séquences se fait une par une ce qui facilite la parallélisation du processus. Les outils de placement phylogénétique tel que pplacer (Matsen et al., 2010), évaluent *a postériori* la probabilité d'un placement sur une branche, permettant ainsi de mesurer l'incertitude sur chaque branche. Cet outil, adapté aux grands jeux de données, peut être lancé en lignes de commandes sur une machine locale. Il est également implémenté dans PhyloAssigner (Vergin et al., 2013).

Parmi les outils automatisés implémentant une démarche phylogénétique complète (incluant l'alignement à des séquences de référence et la reconstruction phylogénétique), il existe STAP (Wu et al., 2008) qui utilise les outils classiquement employés pour la reconstruction phylogénétique (ClustalW (Thompson et al., 1994) pour les alignements et PhyML (Guindon and Gascuel, 2003) pour la génération des arbres) et insère une sé-

13. Les problèmes NP-complet sont les problèmes NP les plus complexes. Un problème est dit NP si la recherche de sa solution nécessite un temps de calcul au moins exponentiel par rapport à la taille des données ($O(k^n)$). Si N= 100 et k=3, il faudra au moins 3^{100} opérations pour résoudre un tel problème.

quence par arbre. Bien que STAP se base sur la phylogénie pour l'affiliation taxonomique, le fait qu'il procède par la génération d'une phylogénie par séquence ne permet pas de faire une comparaison globale de séquences requêtes, et les groupes monophylétiques sont estimés de manière approximative.

Outre la complexité des calculs employés par les méthodes phylogénétiques, la taille des séquences générées par les techniques de séquençage constitue une limitation supplémentaire à leur utilisation. En effet, la qualité de l'inférence phylogénétique dépend de l'information portée par les séquences ; plus l'information est riche et précise, plus l'arbre généré est robuste. D'après Liu et al. (2007), s'il est possible de restituer la taxonomie de courts fragments avec BLAST ou RDP, les approches d'affiliation phylogénétiques demeurent plus sensibles à la taille des séquences. De nombreuses études ont essayé de mesurer la perte de précision suite à l'utilisation de séquences tronquées de l'ADNr 16S par rapport aux séquences complètes (Liu et al., 2007, 2008, Huse et al., 2008, Youssef et al., 2009), et particulièrement l'effet de ces courts fragments sur les assignations taxonomiques et l'estimation de la diversité. Selon Jeraldo et al. (2011), les fragments de 200 pb produisent des phylogénies dont la topologie est significativement différente des phylogénies des séquences complètes, indiquant une perte du signal phylogénétique. Cependant, l'analyse par ces auteurs de différentes régions hypervariables de l'ADNr 16S afin d'évaluer la région qui restitue le mieux l'information biologique et qui peut être utilisée comme un proxy pour la séquence complète, indique que la région v1-v3 restitue le mieux l'information phylogénétique, et permet de générer à 400 pb des phylogénies semblables aux phylogénies des séquences complètes. Avec les longueurs de séquences d'environ 300 - 500 pb générées par pyroséquençage (FLX Titanium) et depuis peu par MiSeq, il est clair que les méthodes phylogénétiques affichent un net avantage par rapport aux autres méthodes d'annotation, et que leur utilisation sur les données de métagénétique permettrait de mieux comprendre l'évolution et l'écologie des micro-organismes.

Table 1.3. Caractéristiques des outils utilisés en écologie microbienne pour l'annotation taxonomique des séquences d'ARNr 16S et 18S.

	Greengenes	RDP-Py	Silva	GAST	Mothur	QIIME	STAP	WATERS	PANGEA	CANGS
Utilisation	Web	Web	Web	Lignes de commandes	Lignes de commandes	Lignes de commandes	Lignes commandes	Lignes de commandes	Lignes de commandes	Lignes de commandes
16S 18S	Oui Non	Oui Non	Oui Oui	Oui Non	Oui Oui	Oui Oui	Oui Oui	Oui Non	Oui Non	Oui Non
Méthode d'affiliation	Simrank	RDP classifier	Align--ement	Align--ement	BLAST, RDP classifier	BLAST, RDP classifier	Phylogénie NN	Phylogénie NN (STAP)	Megablast	BLAST
Alignements	NAST	Infernal	SINA	MUSCLE	NAST	NAST, MUSCLE, Infernal	CLUSTAL	Infernal, CLUSTAL	- - -	- - -
Phlylogénie	Non	Non	Non	Non	Non	Oui	Oui	Oui	Non	Non
Séquences insérées	-	-	-	-	-	Limite FastTree	1	1	-	-
Volume de donnés	500	500000	1000	-	-	-	-	-	-	-
Référence	DeSantis et al. 2006	Cole et al. 2009	Quast et al. 2013	Huse et al. 2008	Schloss et al. 2009	Caporaso et al. 2010	Wu et al. 2008	Hartman et al. 2010	Giongo et al. 2010	Pandey et al. 2010

Bases de référence

Toutes les approches d'affiliation taxonomique décrites précédemment utilisent des bases de référence contenant un ensemble de séquences annotées, contre lesquelles les séquences expérimentales sont comparées, et dont la qualité affecte la précision de l'assignation taxonomique. En effet, la fiabilité des taxonomies inférées dépend de la taille de la base de référence utilisée, de son taux de couverture de la richesse, de la qualité de ses séquences (e.g., absence de chimères [14] ; gestion des introns [15]) ainsi que de l'exactitude de la taxonomie qu'elle propose. Toutes les séquences du gène codant la petite sous unité d'ARNr dans les bases publiques (environ 3.5 millions en juillet 2012 (Quast et al., 2013)) sont répertoriées et accessibles via les bases de données internationales des séquences nucléotidiques (i.e., ENA, Genbank, DDBJ via le INSDC pour International Nucleotide Sequence Databases Collaboration) (Cochrane et al., 2011). Pour les études d'écologie microbienne visant à décrire la diversité à partir de la petite sous-unité d'ARNr, les bases les plus utilisées sont SILVA (Pruesse et al., 2007, Quast et al., 2013) ; Greengenes (DeSantis et al., 2006) et RDP-II (Cole et al., 2007, 2009). Greengenes annote les séquences d'*Archaea* et de bactéries par rapport aux taxonomies de Bergey's et NCBI, qui ne se basant pas sur une phylogénie peuvent contenir des erreurs (Wang et al., 2007). La base SILVA s'étend aux trois domaines de la vie et contient également des séquences de la grande sous-unité de l'ARNr. Elle assigne la taxonomie sur la base d'alignements vérifiés et de phylogénies, et propose parallèlement les taxonomies de NCBI et Greengenes. Ces bases de données proposent plusieurs critères (e.g., la longueur ; les alignements ; la présence d'homoploymères [16]) pour contrôler la qualité des séquences. Bien qu'une taxonomie cohérente, précise, et standardisée soit nécessaire pour une comparaison systématique des communautés, l'augmentation du nombre des séquences dans les bases de données conduit à une accumulation de taxonomies imprécises (e.g., environmental sample ; unclassified) et/ou erronées en l'absence d'une expertise manuelle. Ce type d'erreurs ou d'imprécisions est retrouvé tout aussi bien dans les groupes bactériens et archéens qu'au niveau des euca-

14. Séquence résultant d'une recombinaison de deux séquences distinctes
15. Insertions dans les séquences d'ARNr mais n'apparaissant pas dans la molécule fonctionnelle. Ils génèrent des hétérogénéités au niveau de la taille des séquences
16. Des insertions constituées d'une suite d'un même nucléotide

ryotes. Selon Guillou et al. (2013), jusqu'à 20% des séquences du gène codant pour l'ARNr 18S soumises dans les bases de données publiques ont une taxonomie incomplète ou inexistante. Il faut savoir que la classification et la nomenclature des groupes eucaryotes, et notamment des protistes, ont subi des modifications radicales durant la dernière décennie à la lumière des nouvelles données phylogénétiques et moléculaires (Adl et al., 2005, Marin and Melkonian, 2010, Leliaert et al., 2012, Adl et al., 2012) conduisant à l'apparition de nouveaux groupements, dits "super-groupes" (e.g., Unikontes ; SAR (Burki et al., 2007) ; Archaeplastides). En plus des rangs taxonomiques majeurs existants (Royaume, Embranchement, Classe, Ordre, Famille, Genre, Espèce), plusieurs inter- et sous-niveaux taxonomiques ont été créés, compliquant d'autant plus la comparaison taxonomique et multipliant les erreurs d'annotation. La base dédiée aux séquences de protistes développée par Guillou et al. (2013) (PR^2 pour Protist Ribosomal Reference database), tend à remédier à ce problème en proposant une nouvelle taxonomie vérifiée, cohérente et organisée en des rangs taxonomiques définis couvrant la diversité des eucaryotes. Par ailleurs, dans cette base, la précision de l'assignation taxonomique est liée à la qualité des séquences ; les séquences de faible qualité ayant une taxonomie peu précise.

Enfin, bien que la qualité et la précision des assignations taxonomiques soient étroitement liées à la qualité de la base de référence utilisée, les nouveaux taxons demeurent difficilement identifiables quelle que soit la base de référence. Avec le potentiel du séquençage massif à capturer des espèces jusque-là inconnues, il est clair qu'en construisant les arbres *de novo*, l'approche phylogénétique reste la plus apte à mettre en évidence des nouvelles lignées (Liu et al., 2008).

1.4 Limites de la métagénétique en écologie microbienne

1.4.1 Impact des erreurs de séquençage sur la richesse

Afin de comparer la richesse estimée d'une communauté par rapport à sa réelle richesse et d'évaluer la précision des différentes méthodes utilisées pour la description de la diversité, des communautés de composition taxonomique connue, dites « communautés maquettes » ont été utilisées (Kunin et al., 2010, Quince et al., 2011, Bonder et al., 2012). Les résultats de ces études montrent qu'après clusterisation, le nombre d'OTUs estimées est supérieur au nombre initial d'OTUs présentes dans la communauté maquette. Selon

ces auteurs, cette inflation d'OTUs, conduisant également à une sur-estimation de la richesse et de la diversité, est essentiellement due aux erreurs inhérentes à la technique de séquençage, à l'amplification par PCR, à la formation des chimères et aux méthodes de clusterisation. Alors que les erreurs de séquençage dans le cadre de la métagénomique peuvent être en partie corrigées grâce à l'assemblage et à la profondeur de séquençage, chaque séquence en métagénétique est considérée *a priori* comme un représentant d'un groupe taxonomique donné et contribue à l'augmentation de la richesse. Par ailleurs, bien que le taux d'erreur par base découlant du pyroséquençage soit comparable à celui découlant du séquençage Sanger (Huse et al., 2007), la différence d'échelle dans le nombre de séquences produit augmente considérablement l'impact de ces erreurs sur l'étude de la diversité. Il est donc nécessaire d'établir des filtres et des critères de qualité permettant de distinguer la diversité réelle de la diversité artefactuelle et ainsi réduire l'inflation des OTUs (Quince et al., 2009, Kunin et al., 2010). Ceci est d'autant plus important pour les espèces rares (Sogin et al., 2006, Huber et al., 2007) car, leur réalité est toujours discutée du fait que les OTUs « faux » résultant des erreurs de séquençage sont présentes à des fréquences basses.

Plusieurs stratégies ont été proposées pour définir quelles étaient les séquences de mauvaise qualité dans les données du pyroséquençage (Huse et al., 2007, Kunin et al., 2010). Ces stratégies se basent sur les scores de qualité (Q) des bases nucléotidiques qui reflètent la probabilité de l'inférence correcte de chaque nucléotide (Q= -10 $log_{10}P$ ou P est la probabilité d'erreur. Une probabilité d'erreur de 10^{-3} se traduit par un score de 30). Les filtres les plus utilisés dans la littérature sont d'écarter les séquences contenant des bases indéfinies (Ns) ; celles contenant une ou plusieurs erreurs dans les amorces et les clés qui identifient les échantillons (barcode) ; celles avec un faible score de qualité ; et celles dont la longueur est inférieure à une longueur minimale définie par l'utilisateur (e.g. PyroTagger (Kunin and Hugenholtz, 2010)). Une autre approche utilisée par Zaura et al. (2009), se base sur le fait que des séquences contenant des erreurs ont tendance à former de nouvelles OTUs à une ou deux séquences (singletons ou doublets) quand la distance qui les sépare d'une OTU vraie est plus élevée que le seuil de clusterisation. Ils ont employé par conséquent des seuils plus conservatifs (94% et 90%) et ont éliminé les OTUs avec moins de cinq incidences afin de ne pas surestimer la diversité. L'approche

développée par Huse et al. (2010) appelée « single-linkage preclustering » (SLP), consiste à dérépliquer, c'est à dire à regrouper les séquences strictement identiques pour n'en retenir qu'un unique représentant, et à les classer par ordre décroissant de fréquence. Ensuite, les séquences uniques ou centroïdes recrutent les autres séquences si la distance qui les sépare d'elles est inférieure à un seuil donné. Cette méthode suppose que les vrais positifs sont plus fréquents que leurs variantes contenant des erreurs, et assigne de ce fait les variantes peu fréquentes aux clusters les plus abondants.

Une stratégie différente consiste en la détection et la correction du « bruit » sur les séquences en utilisant les « flowgram [17]» qui reflètent les scores de qualité des nucléotides (e.g. DeNoiser (Reeder and Knight, 2010) ; AmpliconNoise (Quince et al., 2011) ; Pyro-Noise (Quince et al., 2009)). De la même manière que le SLP, ces algorithmes se servent des fréquences des amplicons pour détecter les erreurs ; ils corrigent les séquences qui sont statistiquement plus susceptibles d'être des variantes d'un vrai positif suite à une erreur de séquençage, que correspondre à de nouvelles espèces rares. Implémentant une approche itérative, ces algorithmes sont coûteux en ressources informatiques. Ils se basent sur le regroupement des flowgrams plutôt que des séquences et supposent que, d'une part les séquences contenant des erreurs sont rares, et d'autre part qu'elles doivent être assignées à une « vraie » séquence abondante. De plus, ils ne permettent pas de détecter les erreurs dues à l'amplification par PCR et qui sont invisibles sur les flowgrams. Le programme AmpliconNoise est intégré dans QIIME (Caporaso et al., 2010) et Mothur (Schloss et al., 2009), deux logiciels largement utilisés en écologie microbienne pour le traitement des données de pyroséquençage. Plusieurs études traitant des données du pyroséquençage ont employé ces différentes stratégies de nettoyage afin de comparer leur stringence (Vila-Costa et al., 2012, Comeau et al., 2012). Ces auteurs concluent que les filtres classiques (pas de Ns, pas d'erreurs sur les amorces, choix d'un seuil de qualité et de taille minimal) donnent les mêmes résultats que les méthodes de correction du bruit de séquençage, et qu'ils sont moins coûteux en temps de calcul. Enfin, l'étude de De León et al. (2012) a mis en relation les paramètres à appliquer et la région sur l'ARNr 16S à considérer, et les

17. Flowgram : le signal ou flowgram des séquences indique la nature des nucléotides et leurs scores de qualité sur la base des propriétés relatives aux distributions statistiques du contôle des fragments de l'ADN.

auteurs préconisent d'augmenter le seuil des scores de qualité pour les régions riches en homopolymères.

En plus des erreurs de séquençage, les erreurs engendrées par la PCR, indépendamment des techniques de séquençage, biaisent également l'estimation de la diversité. D'après Lee et al. (2012), les biais de PCR sont plus importants que les erreurs de séquençage. Ils préconisent par ailleurs le développement de méthodes permettant de caractériser l'origine de ces erreurs et de les corriger, étant donné qu'elles ne sont pas traitées par les filtres qualité actuels. De plus, les PCR sont aussi à l'origine de la formation de chimères, séquences artefactuelles obtenues par l'association de deux ou plusieurs produits d'amplification, notamment pour de longues séquences. Ces recombinaisons peuvent avoir lieu à partir de n'importe quelle position sur un amplicon, néanmoins la probabilité de leur formation augmente avec la taille de ce dernier et reste faible sur de courts fragments (Cronn et al., 2002).

Selon la provenance des séquences à l'origine de la formation de la séquence hybride, il existe deux méthodes de détection des chimères : soit si après comparaison de la séquence contre une base de référence, celle-ci présente des similitudes avec deux ou plusieurs séquence de référence, elle est dite chimérique ; soit toutes les séquences expérimentales sont comparées les unes par rapport aux autres, et une séquence est dite chimérique quand elle présente des similitudes avec deux ou plusieurs séquences du jeu de données (détection *de novo*). La méthode de comparaison à une base de référence ne permet de détecter les chimères que si la base contient les séquences parentes. De plus, l'abondance différentielle des groupes taxonomiques dans la base peut résulter en une identification préférentielle des chimères. La détection *de novo* quant à elle, détecte les chimères même quand elles ne possèdent pas de parents connus dans les bases de référence, mais ne doit être utilisée que sur les produits d'une même amplification afin de n'avoir que les parents potentiels. Plusieurs outils adaptés au séquençage massif ont été développés pour l'identification des chimères, e.g., ChimeraSlayer (Haas et al., 2011) ; Perseus (Quince et al., 2011) et UCHIME (Edgar et al., 2011). Pour des amplicons correspondants aux gènes codant l'ARNr 16S, il a été montré que UCHIME et Perseus (détection *de novo*) sont plus performants que ChimeraSlayer (comparaison à une base) (Edgar et al., 2011, Haas et al., 2011).

1.4.2 Impact de la profondeur de séquençage sur les mesures de diversité

En écologie microbienne, les mesures de diversité basées sur l'abondance relative des séquences constituent une approximation de l'abondance relative des organismes dont elles proviennent car intuitivement, les organismes dominants dans une communauté domineraient également en nombre de séquences. Avec le séquençage massif, la capacité à prédire la prévalence d'un organisme dans un environnement à partir de la fréquence des séquences qui lui sont affiliées est biaisée par de nombreux facteurs méthodologiques pour lesquels le changement d'échelle amplifie l'impact. En effet, en plus de la variation du nombre de copies du gène codant la petite sous unité de l'ARNr entre micro-organismes (e.g., le nombre des copies du gène peut atteindre jusqu'à 300 000 copies chez les ciliés (Gong et al., 2013)) qui fausse l'estimation de leur fréquence relative (Rooney and Ward, 2005, Lee et al., 2009), d'autres biais associés à l'extraction de l'ADN (DeSantis et al., 2005, Feinstein et al., 2009) ; aux amorces utilisées (Jumpponen, 2007, Engelbrektson et al., 2010) et à la PCR (Polz and Cavanaugh, 1998) ont été mis en évidence. A ces biais s'ajoutent les biais inhérents au séquençage massif, à savoir i) l'ajout de barcodes en amont des amorces pour différencier les échantillons lors d'un séquençage multiplex avec une amplification préférentielle de certains échantillons (Berry et al., 2011), ii) l'étape d'amplification par PCR en émulsion lors du pyroséquençage et iii) les erreurs de séquençage (Huse et al., 2007, Quinlan et al., 2008, Rozera et al., 2009, Kunin et al., 2010). Ainsi, la combinaison de ces facteurs résulte en une profondeur de séquençage non contrôlée générant des jeux de séquences de taille différente en fonction des échantillons. Par ailleurs, même si cette différence de taille ne déforme pas la réalité biologique des mesures de diversité visant à comparer plusieurs échantillons (béta-diversité) ayant subi les mêmes traitements et donc les mêmes biais (Amend et al., 2010), elle affecte par contre l'estimation de la diversité au sein d'un échantillon (alpha-diversité). Une étude menée par Gihring et al. (2012) a démontré que les indices de richesse non paramétriques Chao1 (Chao, 1984) et ACE (Chao and Lee, 1992) sont hautement sensibles à l'effort de séquençage, ce qui rend toute comparaison multiple basée sur ces indices impossible. Lemos et al. (2011) ont comparé les indices de diversité Simpson (Simpson, 1949) et Shannon (Ludwig and Reynolds, 1988) et ont conclu que l'indice Shannon est moins sensible à la taille des

échantillons. De plus, cette étude a démontré l'avantage des approches de mesure de la diversité basées sur les distances phylogénétiques, notamment l'outil Unifrac (Lozupone and Knight, 2005, Lozupone et al., 2006) qui prend en compte l'abondance relative de chaque échantillon.

Afin de contourner le problème de l'inégalité des jeux des séquences, (Gihring et al., 2012) proposent de normaliser la taille des échantillons en re-échantillonnant le même nombre de séquences dans chacun d'eux pour calculer les indices de richesse et de diversité. Toutefois, l'indice Chao1 tend à sous-estimer la richesse dans les échantillons de faible taille (Colwell and Coddington, 1994). Selon Hughes et al. (2001), cela vient principalement du fait que Chao1 est fortement corrélé à la taille de l'échantillon jusqu'à ce que le nombre d'espèces observées atteigne au moins la racine carrée de deux fois la richesse totale. Son utilisation est inappropriée pour la comparaison d'environnements pour lesquels les jeux de séquences sont de faibles effectifs ou d'effectifs différents. Par ailleurs, la profondeur de séquençage requise pour une meilleure estimation de la diversité dépend des indices utilisés et des jeux de données (Lundin et al., 2012). Une autre méthode pour remédier au biais occasionné par la profondeur du séquençage, est la stratégie employée en métatranscriptomique par Gifford et al. (2010). Cette stratégie consiste en l'introduction d'un standard interne dans les échantillons pour estimer la profondeur de séquençage et calculer un facteur de correction permettant de convertir les abondances relatives en abondances absolues. Ainsi, les indices de richesse et de diversité sont calculés à partir de jeux de séquences dont la différence de taille reflète une réalité écologique.

1.5 Objectifs de l'étude et organisation du mémoire

D'après cette revue bibliographique, il apparaît que les innovations récentes dans les techniques de séquençage ont permis dans le cadre des études en écologie microbienne de passer de l'analyse de quelques centaines de séquences par étude à des centaines de millions de séquences. Cette différence quantitative des données produites a induit des différences qualitatives quant aux études réalisées. En effet, avec le changement du type de données, les approches classiques d'analyse ne peuvent plus être appliquées et il est devenu nécessaire de définir de nouvelles stratégies en tenant compte des contraintes que

posent ces données. Alors qu'il était possible d'insérer classiquement quelques dizaines de séquences issues des techniques de première génération dans des phylogénies expertisées en utilisant des outils tels que ClustalW (Thompson et al., 1994) et PhyML (Guindon and Gascuel, 2003) ou le package Arb (Ludwig et al., 2004), le nombre de séquences générées aujourd'hui par les NGS à chaque expérience rend cette tâche irréalisable et nécessite la mise en place de nouvelles stratégies et l'utilisation d'outils adaptés. Par ailleurs, les outils d'analyse de la diversité microbienne adaptés aux amplicons de nouvelle génération (e.g., Mothur (Schloss et al., 2009), QIIME (Caporaso et al., 2010)), PyroTagger (Kunin and Hugenholtz, 2010) implémentent des approches probabilistes et/ou de recherche de similitude pour l'identification des séquences environnementales. L'approche phylogénétique quant à elle, bien qu'elle soit la plus robuste, n'est pas utilisée pour l'annotation taxonomique de ce type de données du fait de ses besoins en temps et en ressources de calcul. Au-delà de l'approche d'annotation taxonomique, les nouvelles techniques de séquençage posent également le problème de la qualité des séquences produites et son impact sur l'estimation de la diversité.

Ainsi, l'objectif de cette thèse a été de définir une stratégie d'analyse bio-informatique de données de séquençage massif dans le contexte de l'étude de la diversité microbienne, en tenant compte des limitations imposées par les ressources informatiques actuelles (matérielles et logicielles) d'un côté, et de l'avantage des méthodes phylogénétiques par rapport aux autres approches d'annotation taxonomique. Ce travail a donné lieu au développement d'une chaîne de traitement proposant une série d'analyses allant des séquences brutes jusqu'à la visualisation des résultats, tout en replaçant les séquences environnementales dans un contexte évolutif.

Le mémoire de thèse s'articule donc autour des points suivants :

– Le chapitre II présente l'implémentation d'une procédure d'annotation phylogénétique automatisée pour l'affiliation taxonomique et la description de groupements monophylétiques. L'approche développée a été optimisée pour la gestion de gros volumes de données, et a été comparée en terme de précision d'affiliation aux autres approches communément utilisées en écologie microbienne (BLAST et RDP Clas-

sifier). Les tests et simulations ont montré qu'à partir d'une taille d'amplicons de 400 pb, l'affiliation phylogénétique avait les meilleurs résultats mais aussi, que la qualité de cette affiliation différait selon la région hypervariable ciblée. La chaîne de traitements mise en place a ensuite été implémentée dans un contexte de calcul à haute performance, notamment sur un cluster de calcul, pour proposer un service web dédié à l'analyse de la diversité microbienne.

– Le chapitre III décrit l'application de la méthode phylogénétique à des données de pyroséquençage obtenues dans le cadre de deux études sur la dynamique des communautés microbiennes provenant de deux écosystèmes différents. La première étude sur la diversité et l'activité des communautés des *Archaea* en milieu marin montre que la profondeur de séquençage actuelle couplée à une approche phylogénétique permet de détecter la fraction rare et de mettre en évidence de nouveaux clades. Dans la deuxième étude, l'ajout d'un standard interne aux différents échantillons d'ADN, en quantités définies avant PCR et séquençage, montre qu'il est possible de quantifier les erreurs sur les séquences et de choisir le seuil de clusterisation le plus adapté à l'expérience. Il nous a également permis d'évaluer l'effet des différences en terme de profondeur de séquençage entre les différents échantillons et de normaliser leurs effectifs.

– Enfin, le chapitre IV aborde, sous la forme d'une discussion générale, l'ensemble des résultats obtenus. Il vise à discuter i) des différents aspects méthodologiques mis à disposition pour l'étude de la diversité microbienne à la lumière de la métagénétique à haut débit, ii) à identifier les limites encore non résolues pour les différentes approches proposées et iii) à proposer des perspectives pour améliorer les traitements et l'affiliation des données de NGS.

Chapitre 2

Méthodologie

2.1 Introduction

L'approche phylogénétique est considérée comme supérieure aux autres approches d'annotation taxonomique car elle réévalue l'histoire évolutive des séquences et les liens de parenté entre organismes (Price et al., 2009). Alors que l'utilisation de cette approche était initialement limitée pour le traitement des données issues des NGS du fait de la taille réduite des fragments générés et donc de leur faible signal phylogénétique, les progrès dans les technologies de séquençage produisent actuellement des fragments plus longs offrant une meilleure résolution phylogénétique (Figure 2.1). D'après les simulations que nous avons réalisées, il apparaît que le taille actuelle des séquences (≈450 pb) permet une restitution phylogénétique d'environ 60% au niveau du genre, engendrant un gain en précision de 30% par rapport aux fragments de 250 pb. Ces résultats sont en accord avec les conclusions de Jeraldo et al. (2011) selon lesquelles les fragments d'environ 400 pb restituaient mieux l'information biologique.

Toutefois, ces méthodes demeurent peu utilisées pour les données du séquençage massif car elles sont difficiles à automatiser, et ce pour deux raisons principales : i) la fiabilité des résultats d'affiliations phylogénétiques dépend de la qualité des alignements qui doit être vérifiée et validée ; ii) ces méthodes sont très coûteuses en temps de calcul, les étapes d'alignement et de construction phylogénétique étant les plus longues et nécessitant beaucoup de ressources informatiques. En effet, une assignation taxonomique par une approche phylogénétique consiste à :

- Construire une base de référence contre laquelle les séquences expérimentales seront comparées ;
- Aligner les séquences expérimentales avec celles contenues dans cette base ;
- Construire l'arbre phylogénétique à partir de l'alignement ;
- Parcourir l'arbre résultant pour identfier la position des séquences insérées ;

Parmi les outils d'analyse phylogénétique automatisée actuellement disponibles, BIBI (Devulder et al., 2003) permet l'annotation des séquences bactériennes. Il utilise BLAST pour l'identification des séquences similaires à la séquence requête et ClustalW (Thompson et al., 1994) pour générer l'alignement multiple. STAP (Wu et al., 2008), quant à

lui, couvre les trois domaines de la vie. Il utilise BLAST pour déterminer l'ensemble des séquences homologues à la séquence requête et ClustalW pour le calcul de l'alignement multiple, et construit l'arbre phylogénétique avec PhyML (Guindon and Gascuel, 2003). Les arbres sont ensuite parcourus par un script et les assignations taxonomiques sont inférées selon la taxonomie du voisin le plus proche (Nearest Neighbor). La base de données utilisée par STAP regroupe des séquences du gène de l'ARNr 16S des bactéries et des *Archaea* extraites de Greengenes (DeSantis et al., 2006), et celles d'ARNr 18S de RDP II (Cole et al., 2007). Même si STAP est complètement automatisé, les outils qu'il implémente ne sont pas adaptés à de grands jeux de données. De plus, il génère un arbre par séquence expérimentale ce qui empêche la détection de nouveaux clades potentiels.

Ces deux outils se basent sur le principe de la réduction de l'espace des données, ils souscrivent à partir de toutes les séquences présentes dans la base de données un sous ensemble de séquences bien annotées qui reflète la diversité des groupes taxonomiques représentés dans les données expérimentales pour accélérer les traitements phylogénétiques. Par ailleurs, il est nécessaire de s'assurer que la perte d'information inhérente à la réduction de l'espace des données n'a pas de conséquence sur la représentativité de la richesse dans la base ainsi construite.

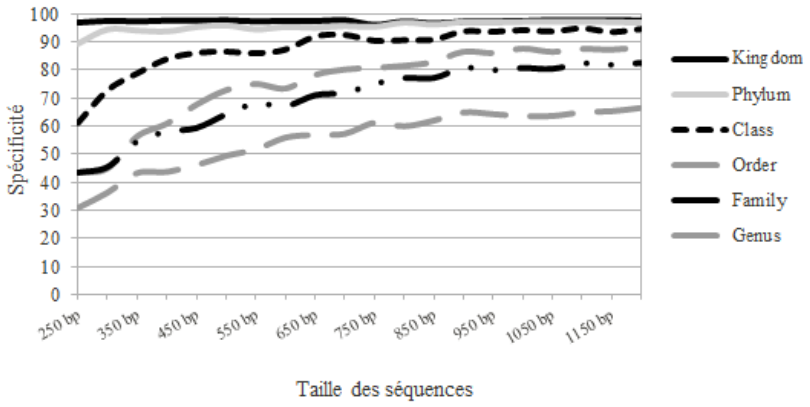


Figure 2.1. Représentation graphique de la spécificité de l'affiliation phylogénétique en fonction de la taille des séquences et du rang taxonomique.

La spécificité a été calculée à partir d'un jeu de données comprenant 1000 séquences ; la taille des séquences correspond au nombre de bases compté à partir de la première base.

Dans l'étude présentée ci-après, nous avons développé une procédure d'annotation phylogénétique automatisée et adaptée à de grands jeux de données, en se basant sur des

outils d'alignement et de construction phylogénétique disponibles. Pour cela, en amont du travail publié, plusieurs outils ont été testés et comparés en termes de temps de traitement, de besoins en ressources informatiques et de précision. Une base de référence a également été construite à partir des séquences alignées de qualité issues de SILVA (Pruesse et al., 2007, Quast et al., 2013).

2.2 Choix méthodologiques

2.2.1 Alignement

Parmi les approches d'alignement de séquences disponibles, on distingue les alignements *de novo* de ceux contre des profils. Un alignement de profil permet d'insérer une ou plusieurs séquences dans un alignement multiple de référence. Un profil correspond à une matrice où les colonnes sont des vecteurs de probabilité qui dénotent la fréquence de chaque nucléotide dans la colonne d'alignement correspondante. Cette approche a été retenue afin d'utiliser des alignements dont la qualité a fait l'objet d'une expertise manuelle, en l'occurrence ceux de la base de données SILVA.

Quatre outils d'alignement ont été comparés en terme du temps nécessaire pour insérer une séquence au sein d'alignements de différentes tailles. D'après les résultats obtenus, il apparaît que les outils disponibles sont encore limités pour une utilisation sur des machines locales, et ne peuvent être appliqués sur les sorties des nouvelles techniques de séquençage (des millions de séquences). En revanche, pour des jeux de données de taille modérée, hmmalign (Eddy, 1998) semble le plus adapté en terme de temps de traitement (tableau 2.1).

Étant donné que les alignements globaux présentent l'inconvénient d'étaler les séquences courtes le long de l'alignement, générant des brèches au niveau des colonnes de l'alignement (par exemple lors d'une insertion spécifique dans une séquence d'un groupe taxonomique donné) et une perte du signal, les profils construits à partir de ces alignements sont peu efficaces et altèrent la qualité de l'alignement. Afin d'augmenter la qualité des alignements issus de profils et la qualité du signal dans chaque profil, nous avons implémenté une approche permettant d'une part de restreindre les profils à des groupes taxonomiques monophylétiques, et d'autre part de limiter chaque profil à la région ciblée par les amplicons.

Table 2.1. Comparaison des temps de traitement de différents outils d'alignement pour l'insertion d'une séquence dans deux profils de taille différente.

Nombre de séquences x Taille de l'alignement	Muscle	ClustalW	Multalin	hmmalign
6 328 x 50 000	3h 48min	-	1min 27s	< 1s
6 328 x 21 807	2h	1h 17min	1min 27s	< 1s

Les alignements sont réalisés contre deux profils différents avec les outils Muscle (Edgar, 2004) ; ClustalW (Larkin et al., 2007) ; Multalin (Corpet, 1988) et hmmalign (Eddy, 1998). Les cases vides indiquent que l'alignement n'a pas pu être calculé.

2.2.2 Phylogénie

Parmi les approches de reconstruction phylogénétique, les méthodes basées sur le maximum de vraisemblance (e.g., PhyML (Guindon and Gascuel, 2003)) sont les plus robustes quand les données s'écartent des hypothèses du modèle, mais elles sont lentes en terme de temps de calcul. Celles basées sur la distance comme le neighbor joining (e.g., BIONJ (Gascuel, 1997)) sont moins robustes mais produisent des arbres satisfaisants quand les données ne s'écartent pas trop du modèle (Graur and Li, 2000).

Dans le contexte du séquençage massif, il existe deux outils pour la reconstruction phylogénétique, RAxML (Stamatakis, 2006, Stamatakis et al., 2008) et FastTree (Price et al., 2009, 2010). RAxML implémente une approche de maximum de vraisemblance, alors que FastTree implémente une approche heuristique du maximum de vraisemblance en interchangeant les séquences voisines sur un nœud de l'arbre au lieu de vérifier toutes les possibilités, réduisant ainsi considérablement le temps de calcul. Ainsi, à partir de l'alignement de 6328 séquences de 21807 positions, FastTree génère une phylogénie en 48 minutes. Selon une étude réalisée par Liu et al. (2011) comparant RAxML et FastTree, il apparaît que ces deux outils génèrent des topologies équivalentes, FastTree étant 100 à 1000 fois plus rapide. Ainsi, même si RAxML présente généralement de meilleurs scores de vraisemblance, son utilisation reste limitée par son temps de calcul.

Afin d'évaluer la fiabilité de l'approche heuristique implémentée dans FastTree, des phylogénies de plusieurs groupes monophylétiques ont été extraites d'Arb (Ludwig et al., 2004) accompagnés des alignements SILVA ayant servi à les construire avant d'être à nouveau générées à l'aide de trois outils : PhyML implémentant une approche de maximum

Table 2.2. Comparaison des topologies des phylogénies générées par différents outils par rapport au maximum de vraisemblance.

Phylogénie	PhyML	BIONJ	FastTree
Parabasilidea	**1.000**+	0.0000	**0.1970**+
Litostomatea	**1.000**+	0.0380	**0.3690**+
Haptophyceae	**1.000**+	0.0000	0.0040
Foraminifera	**1.000**+	0.0000	**0.4750**+
Euglenozoa	**0.7510**+	0.0000	**1.000**+
Cryptophyta	**0.6840**+	0.0290	**1.000**+
Apicomplexa	**1.000**+	0.0010	**0.3830**+
Ciliophora	**1.000**+	0.0000	0.0560
Haemosporida	**1.000**+	**0.3950**+	**0.7710**+

Les résultats du test de Shimodaira-Hasegawa du maximum de vraisemblance pour comparer les topologies des phylogénies générées par PhyML, BIONJ et FastTree avec la meilleure phylogénie du maximum de vraisemblance, pour différents groupes monophylétiques eucaryotes.

de vraisemblance ; BIONJ implémentant une approche neighbor joining et FastTree. Les topologies des trois phylogénies ont ensuite été comparées à celle générée par le maximum de vraisemblance avec le test Shimodaira-Hasegawa (Shimodaira and Hasegawa, 1999) tel que implémenté dans le package tree-puzzle (Schmidt et al., 2002). Comme illustré dans le tableau 2.2, alors que les phylogénies générées par BIONJ possèdent les scores les plus faibles, traduisant ainsi l'éloignement de leurs topologies par rapport au maximum de vraisemblance (PhyML). FastTree quant à lui, même avec des scores plus faibles que ceux de PhyML, génère des phylogénies dont la vraisemblance n'est pas significativement différente de celles produites au maximum de vraisemblance, à l'exception des groupes *Haptophyceae* et *Ciliophora*.

2.2.3 Recherche de similitude

La première étape de la procédure de l'affiliation taxonomique des séquences expérimentales développée ici est basée sur une comparaison des séquences expérimentales avec une base de référence afin de trier les séquences selon les profils taxonomiques. Cette comparaison est réalisée par une approche de recherche de similitude afin d'attribuer aux séquences une première affiliation. De cette affiliation, seul le phylum sera retenu et utilisé pour trier les séquences expérimentales afin qu'elles soient redistribuées sur le profil d'alignement correspondant. Pour cette première affiliation, nous avons testé deux outils :

BLAST (Altschul et al., 1997) et UCLUST (Edgar, 2010).

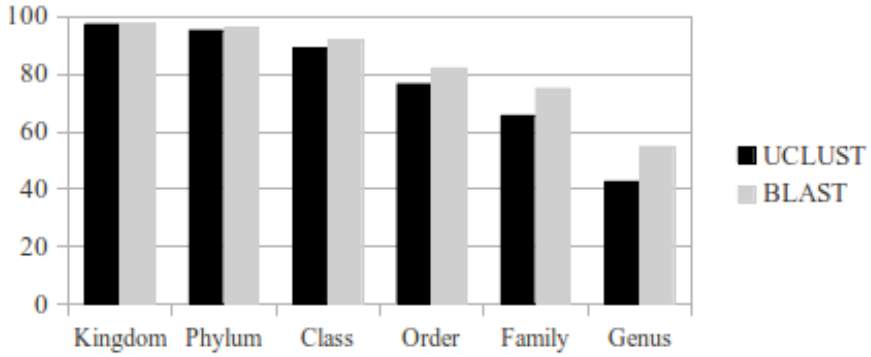


Figure 2.2. Histogramme représentant la spécificité moyenne de l'affiliation de BLAST et UCLUST calculée sur 4 x 1000 séquences complètes.

La figure 2.2 représente la spécificité (les affiliations correctes par rapport aux taxonomies initiales) de BLAST et UCLUST par niveau taxonomique calculée sur 4 x 1000 séquences complètes retirées de la base de référence. Il apparaît de cette figure que, même si le BLAST donne de meilleurs résultats à de bas niveaux taxonomiques (Ordre, Famille, Genre), UCLUST a une spécificité équivalente au rang du phylum. De plus, sur une machine avec une CPU Intel(R) Xeon(R) 2 GHz et 24 GB de RAM, UCLUST compare 1000 séquences complètes à une base de 21100 séquences en moins d'une minute alors que BLAST réalise cette comparaison en 75 minutes. UCLUST peut donc être utilisé pour une affiliation rapide à un niveau phylogénétique peu résolutif.

Parallèlement aux aspects de vitesse et de besoins en ressources informatiques, les outils retenus (hmmalign, FastTree et UCLUST) ont également été testés en terme de fiabilité des résultats obtenus. En effet, les annotations phylogénétiques inférées à partir de ces outils sur des séquences de référence ont été évaluées en comparant la taxonomie retrouvée à la taxonomie initiale des séquences, et comparées aux affiliations d'autres approches (UCLUST et RDP Classifier). Ces outils ont été intégrés dans un pipeline de traitement de données du pyroséquençage nommé PANAM, présenté dans l'article 1.

2.3 Calcul distribué et parallélisation

Au-delà du choix des outils pour optimiser le temps de traitement des données NGS, l'utilisation de structures de calcul distribué (grille de calcul ou cluster) permet de réduire considérablement les délais de traitement de ces données. En effet, ces infrastructures proposent des ressources informatiques de calcul et/ou de stockage qui, avec des protocoles et des interfaces standardisés, offrent des qualités de service non triviales. Elles permettent ainsi de mobiliser les ressources de plusieurs ordinateurs au sein d'un réseau sur un seul problème nécessitant un grand nombre de cycles de calcul, ou l'accès à de grandes quantités de données. Ces approches passent par la parallélisation des processus lorsque le traitement des données implique de nombreux calculs indépendants, ou par la parallélisation des données lorsque le processus est divisé en plusieurs sous tâches qui peuvent être traitées indépendamment. Dans le cadre de notre étude, nous avons réalisé une parallélisation par les données.

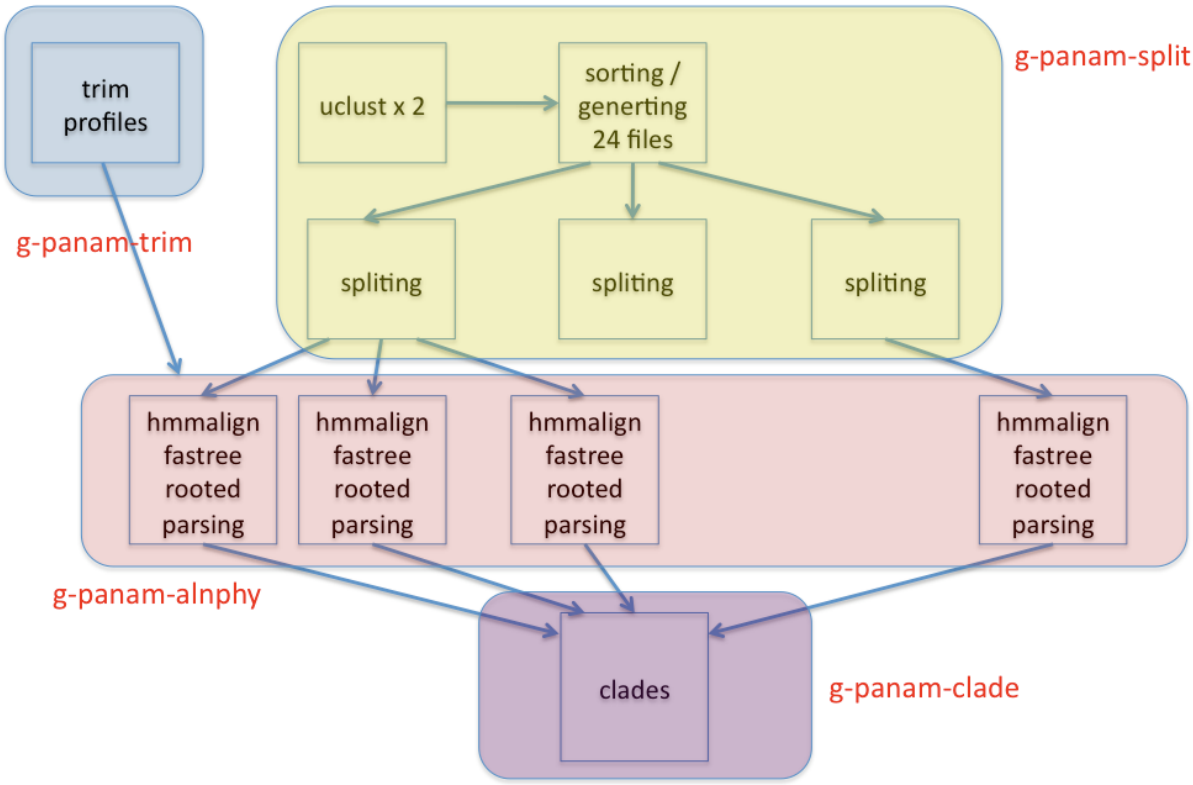
2.3.1 Grille de calcul

Dans le but d'évaluer les performances de PANAM sur la grille de calcul, des développements pour le portage de PANAM sur grille ont été réalisés en collaboration avec l'équipe Plateforme de Calcul pour les Sciences du Vivant (PCSV) du Laboratoire de Physique Corpusculaire (LPC – UMR6533), dans le cadre de la thèse de Doan Trung Tung (Doan, 2012). Comme dit précédemment, la grille est une infrastructure réunissant des moyens de calcul (CE - Compute Element) et de stockage (SE – Storage Element) hébergés et maintenus dans de multiples centres distincts. Ces moyens sont mis en commun et mutualisés grâce à une couche logicielle (intergiciel ou middleware) permettant un accès «transparent» des utilisateurs aux ressources distantes. Le déploiement de PANAM sur la grille a été basé sur le logiciel « Wisdom Production Environment » (WPE, (Breton et al., 2009), installé sur différents éléments de calcul pour gérer les données et les jobs (les programmes avec leurs paramètres et entrées), et pour partager la charge sur l'ensemble des ressources intégrées, même si elles utilisent des technologies différentes. En se basant sur ce WPE, PANAM a été divisé en quatre services (Figure 2.3) :

1. Le service g-panam-trim qui adapte les profils d'alignement des séquences de référence à la région amplifiée des séquences expérimentales avant de lancer le traitement

phylogénétique.

2. Le service g-panam-split trie les séquences selon le groupe phylétique inféré par UCLUST et les affecte aux différents groupes phylétiques.
3. Le service g-panam-alnphy aligne des séquences par hmmalign, avant de construire les phylogénies par FastTree. Cette phase est parallélisée selon le nombre de fichiers récupérés à la fin du service précédant (g-panam-split).
4. Finalement, le service g-panam-clade parcourt les arbres, assigne une taxonomie pour chaque séquence et infère les clades putatifs.



D'après Doan (2012)

Figure 2.3. Représentation graphique du découpage des tâches de PANAM en 4 services WPE.

Ainsi, alors que PANAM sur un machine personnelle réalise successivement chaque alignement de profil et chaque phylogénie, la version parallélisée permet de lancer en même temps autant d'alignement, puis autant de phylogénie qu'il y a de fichiers de données produits. Les simulations sur la grille ont été réalisées sur des fragments de 400 pb. Le tableau 2.3 montre les temps moyens de traitement nécessaires à chacun des trois services ainsi que la durée totale du processus en fonction de jeux de données de taille différente

Table 2.3. Temps moyens d'exécution des différents services de PANAM sur la grille pour différents jeux de séquences

Workflow	Total	g-panam-split	g-panam-alnphy	g-panam-clade
Test107s250k	5h32	3h15	0h20	2h17
Test88s500k	6h22	2h33	0h22	3h48
T53s750k	9h23	8h6	0h23	0h57
T81s1M	12h43	7h18	0h27	0h25

D'après Doan (2012)

en plusieurs répétitions. Pour un jeu de 1 million de séquences de 400 pb, le temps de traitement total nécessaire à PANAM après déploiement sur la grille est 12h 43min, contre 16 jours pour sa version non parallélisée.

Ces travaux montrent l'intérêt des structures de calcul distribué pour de grands jeux de données, confirmant l'efficacité de la grille pour le traitement de données issues des NGS. Ils ont également permis de pointer un certain nombre de points limitants, notamment à cause de la dernière étape de PANAM (clade) qui implique la fusion de l'ensemble des résultats précédemment obtenus. Ainsi, le traitement d'un jeu de données par PANAM dans l'environnement WPE sera conditionné par la durée de la tâche la plus longue. Si pour une raison ou une autre l'une des tâches ne se termine pas, le traitement du jeu de données peut en théorie durer un temps infini. Ces observations ont permis de proposer des modifications du WPE qui en ont considérablement amélioré les performances (Doan, 2012).

2.3.2 Cluster de calcul

Parallèlement à la grille, des développements ont également été faits pour déployer PANAM sur cluster. Contrairement à la grille, les clusters de calcul possèdent une structure plus homogène et sont localisés sur un seul site. L'objectif de ces développements était de pouvoir à terme rendre les traitements de PANAM accessibles à la communauté des microbiologistes à travers une interface graphique (figure 2.4.a) et supportés par des ressources de calcul adaptées. Ici, le traitement initial de PANAM a été découpé en douze sous tâches. De même que pour la grille de calcul, les étapes parallélisées sont celles de l'alignement et de la phylogénie. Un gestionnaire de workflow a été mis en place pour gérer la soumission des données à traiter par le cluster. Les transferts de données à traiter

depuis le serveur web jusqu'au nœud maître du cluster, ainsi que l'envoi des résultats du nœud vers le serveur web sont également gérés par le gestionnaire de workflow.

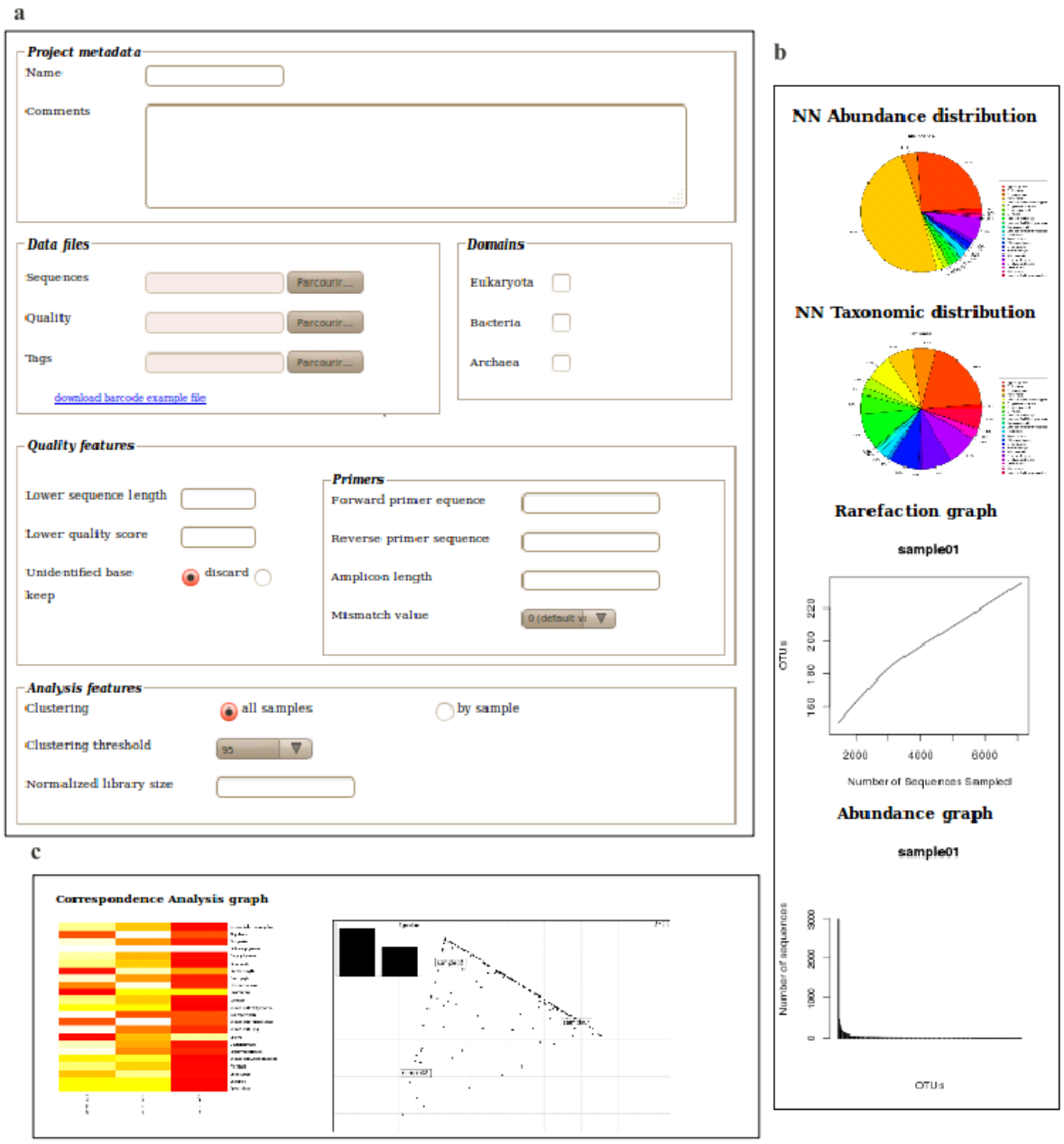


FIGURE 2.4. Captures d'écran de ePANAM, correspondant au formulaire de saisie (a), les graphes générés pour l'alpha-diversité (b) et la béta-diversité (c).

De ces travaux sur le cluster a découlé la mise en place d'un service web ePANAM (présenté dans l'article 2), dédié au traitement des données issues des pyroséquenceurs par une approche phylogénétique, et optimisé pour permettre une analyse complète et

automatisée de grands jeux de données. ePANAM affilie la taxonomie à partir des phylogénies réalisées sur un million de séquences en se basant sur les scripts de PANAM. Les résultats générés sont enfin exploités afin de proposer une description des environnements étudiés par le calcul d'indices de diversité et de richesse (alpha diversité) (figure 2.4.b) ainsi que leur comparaison (béta diversité) (figure 2.4.c).

Article 1

Phylogenetic Affiliation of SSU rRNA Genes Generated by Massively Parallel Sequencing : New Insights into the Freshwater Protist Diversity

OPEN ACCESS Freely available online

PLOS ONE

Phylogenetic Affiliation of SSU rRNA Genes Generated by Massively Parallel Sequencing: New Insights into the Freshwater Protist Diversity

Najwa Taib[1,2], Jean-François Mangot[1,2,3,4], Isabelle Domaizon[3,4], Gisèle Bronner[1,2], Didier Debroas[1,2]*

1 Clermont Université, Université Blaise-Pascal, Laboratoire "Microorganismes: Génome et Environnement", BP 10448, Clermont-Ferrand, France, **2** CNRS, UMR 6023, LMGE, Aubiere, France, **3** INRA, UMR 42 CARRTEL, Thonon les Bains, France, **4** Université de Savoie, UMR 42 CARRTEL, Le Bourget du Lac, France

Abstract

Recent advances in next-generation sequencing (NGS) technologies spur progress in determining the microbial diversity in various ecosystems by highlighting, for example, the rare biosphere. Currently, high-throughput pyrotag sequencing of PCR-amplified SSU rRNA gene regions is mainly used to characterize bacterial and archaeal communities, and rarely to characterize protist communities. In addition, although taxonomic assessment through phylogeny is considered as the most robust approach, similarity and probabilistic approaches remain the most commonly used for taxonomic affiliation. In a first part of this work, a tree-based method was compared with different approaches of taxonomic affiliation (BLAST and RDP) of 18S rRNA gene sequences and was shown to be the most accurate for near full-length sequences and for 400 bp amplicons, with the exception of amplicons covering the V5-V6 region. Secondly, the applicability of this method was tested by running a full scale test using an original pyrosequencing dataset of 18S rRNA genes of small lacustrine protists (0.2–5 μm) from eight freshwater ecosystems. Our results revealed that i) fewer than 5% of the operational taxonomic units (OTUs) identified through clustering and phylogenetic affiliation had been previously detected in lakes, based on comparison to sequence in public databases; ii) the sequencing depth provided by the NGS coupled with a phylogenetic approach allowed to shed light on clades of freshwater protists rarely or never detected with classical molecular ecology approaches; and iii) phylogenetic methods are more robust in describing the structuring of under-studied or highly divergent populations. More precisely, new putative clades belonging to Mamiellophyceae, Foraminifera, Dictyochophyceae and Euglenida were detected. Beyond the study of protists, these results illustrate that the tree-based approach for NGS based diversity characterization allows an in-depth description of microbial communities including taxonomic profiling, community structuring and the description of clades of any microorganisms (protists, Bacteria and Archaea).

Citation: Taib N, Mangot J-F, Domaizon I, Bronner G, Debroas D (2013) Phylogenetic Affiliation of SSU rRNA Genes Generated by Massively Parallel Sequencing: New Insights into the Freshwater Protist Diversity. PLoS ONE 8(3): e58950. doi:10.1371/journal.pone.0058950

Editor: Stefan J. Green, University of Illinois at Chicago, United States of America

Received July 13, 2012; **Accepted** February 11, 2013; **Published** March 14, 2013

Funding: Funding came from Le sonseil régional d'auvergne "http://www.conseil-general.com/conseil-regional/conseil-regional-auvergne.htm". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Didier.Debroas@univ-bpclermont.fr

Introduction

The development of molecular ecology was prompted by indisputable evidence that, for most environments on Earth, the majority of existing organisms had not yet been cultured. This evidence came from the analysis of sequences recovered directly from environmental samples. Vast new lineages of microbial life were uncovered by this approach, changing our picture of the microbial world and yielding a phylogenetic description of community membership [1,2]. More precisely, the sequencing of the small sub-unit (SSU) rRNA genes highlighted new monophyletic groups or clades in the environment, such as SAR11 [3] or MGI [4] among the Bacteria and Archaea respectively. Similarly, several new lineages of protists have been discovered in oceanic systems during the last decade [5]. Recent studies conducted in lakes have also highlighted numerous phylogenetic groups, especially putative parasites (Fungi and Perkinsozoa), and this finding is modifying our view of the microbial loop and therefore, the functioning of aquatic ecosystems [6,7].

Recent advances in next-generation sequencing (NGS) technologies are spurring progress in determining the microbial diversity of various ecosystems by highlighting, for example, the rare biosphere and the activity of these low abundance organisms [8,9]. Currently, the pyrosequencing of amplified SSU rRNA gene variable regions is mainly used to determine bacterial and archaeal diversity and structure in various ecosystems, such as soil [10], ocean [11] or gut microbiota [12]. The recent results obtained regarding the composition and structure of the microeukaryote communities using high-throughput amplicon sequencing performed with the Roche 454 pyrosequencing platform in freshwater systems [13,14] have fuelled the current debates on the biogeography of these microorganisms and on the role of the rare biosphere. The taxonomic assignment of such data is often inferred from supervised classification with the Ribosomal Database Project Classifier (RDP) [15], sequence similarity with BLAST [16–18] or both [19,20]. Pairwise identity scores via BLAST remain the most commonly used tool for large eukaryotic datasets [14,21–26]. However, as claimed by Bik et al. [26], assigning accurate taxonomy to eukaryotic operational taxonomic

units (OTUs) is more difficult than the approaches used for Bacteria; the relative paucity of sequences in public eukaryotic databases results in many sequences without significant top BLAST matches [26]. Furthermore, the best BLAST match assigns a single organism as the most likely phylogenetic neighbor, without specifying the level of relatedness (class, order or phylum) of the compared sequences [27].

Phylogenetic methods assess relatedness among various groups of sequences by inserting unknown OTU sequences within a known phylogeny. On the one hand, these methods allow query sequences to be affiliated with their relatives. Tree-based assignment is, therefore in theory, a more robust approach [28] and current FLX Titanium longer reads now make it possible to extract phylogenetic information with a high degree of reliability [29]. On the other hand, phylogenetic analyses allow for the description of clades, which may lead to new insights into the structure and functioning of ecosystems, as previously mentioned. Moreover, these phylogenetic analyses are not limited to the taxonomic assignment of an individual sequence as implemented in bioinformatic pipelines dedicated to NGS and used in microbial ecology studies (mainly on 16S rRNA gene amplicons): phylogenies can also be used to compare environments (beta-diversity) using methods based on tree topology and/or branch length such as the popular tool UNIFRAC (unique fraction metric) [30]. Although more robust, these methods are less frequently used than BLAST or probabilistic classifiers, as they require more computing resources (Table S1). Though large computational capacity is now more accessible (e.g., QIIME [20] can be implemented on a cloud), massively parallel sequencing projects that seek to elucidate the phylogenetic structure of microbial populations are still faced with the attendant computational challenges of classifying the sequences obtained.

In this work, we introduce a tree-based treatment designed for analyzing massively parallel sequencing outputs that automatically affiliates sequences from SSU rRNA gene amplicons and builds phylogenetic trees composed of very large numbers of sequences. As short-read sequence data (e.g., 100 base sequences generated by the Illumina sequencing platform) provide limited phylogenetic resolution [29], our work is focused on the treatment of moderately long ($\sim$ 450 bp obtained for example with Titanium platforms) to near full-length sequences. Designed for the analysis of any microorganism (protists, Bacteria and Archaea), the value of this treatment is highlighted here on the protist diversity as the pipelines dedicated to the study of eukaryotic pyrotags are still scarce. Indeed, 16S rRNA gene reads were widely investigated in previous studies [31–33] to assess bacterial diversity, which enhanced the development of specific 16S rRNA gene analytical tools. However, 18S rRNA gene surveys and tools allowing for the accurate and rapid taxonomic affiliation of protists from NGS data are needed because the number of studies dealing with protists diversity is currently increasing (e.g., [14,34]). We first tested the accuracy and speed of phylogenetic affiliation on large fragments of well-annotated 18S rRNA gene sequences (>1,200 bp) and on short sequences that simulate pyrosequencing outputs. Secondly, the different methods of taxonomic assignment (i.e., tree-based, similarity and probabilistic approaches) were compared with each other, in a first attempt to determine the best method for affiliating protists in the context of massively parallel sequencing of amplicons. Thirdly, the accuracy of phylogenetic affiliation was compared on amplicons covering different variable regions (V1 to V9), and finally, a dataset of original pyrosequencing data obtained from lacustrine small protists was analyzed by the tree-based approach that was developed.

Results

Evaluation of performance on reference sequences

In the analysis of near full-length reference sequences of 18S rRNA gene, taxonomic groups were found in similar proportions to those initially present in the samples. Our phylogenetic affiliation method, referred to as PANAM (Phylogenetic Analysis of Next-generation AMplicons), was more accurate using LCA (lowest common ancestor) assignment for the different taxonomic ranks, ranging from 99.1% to 90.8% versus 98.6% to 86.7% for PANAM using the NN (nearest neighbor) method (Figure 1.A). For comparison, when refining affiliations from kingdom to genus, the accuracy of the standard phylogenetic affiliation using ClustalW [35] and PHYML [36] as implemented in STAP, ranged between 96.1% and 74.6%. At the finest phylogenetic level studied (i.e., genus), BLAST and RDP allowed for the affiliation of 62.3% and 68.4% of reference sequences. Thus, our phylogenetic affiliation method outperformed the other methods on near full-length sequences. However, as environmental sequences are generally quite divergent from referenced ones and their affiliation needs to be checked manually, sequences belonging to freshwater clades [6,7] were also processed by our phylogenetic affiliation method to evaluate how it behaved on these datasets. The phylogenetic analysis of these environmental sequences (Sanger, >1,200 bp) enabled us to retrieve the affiliations obtained by other authors together with the delineation of freshwater clades corresponding to Cercozoa clade [6] and Perkinsea clades 1 and 2 [7] (Figure S1).

Different 18S rRNA gene regions were targeted by simulating amplicons with lengths of 200 and 400 bp starting from a conserved region given by the following forward primers: NSF4, NSF370, NSF573 NSF963, NSF1179 and NSF1419 (Table S2). Because the V8–V9 region is often missing in public databases, the results obtained from this region were based only on 300 sequences included in the reference database. The affiliation results at the genus rank differed according to length, variability within the studied region and method used for taxonomic affiliation (Table 1). For the six regions tested, the accuracy increased with amplicon length for both affiliation methods implemented in PANAM, LCA and NN. Considering the affiliation methods, LCA specificity was higher than that of NN for fragments of 200 bp only for the V1 and V8 regions, and LCA specificity was always better for fragments of 400 bp. The comparisons with the other affiliation tools implemented in pipelines dedicated to pyrosequencing results showed that at 200 bp, BLAST outperformed RDP, STAP and PANAM, with the exception of the V8 region, for which PANAM (LCA) gave the highest result (68.7%). In contrast, for 400 bp amplicons, the most accurate affiliations were obtained with PANAM, with the exception of the V5–V6 amplicon. In this last region, we observed a decrease in the accuracy of the affiliation, coupled with a sharper decline for the phylogeny-based affiliations. The specificity therefore varied between 64.2% (V5–V6) and 79.2% (V8–V9) at the genus level.

In addition to the accuracy of assignment, this phylogenetic affiliation method was developed to optimize processing time for large datasets. Thus with a 2 GHz Intel(R) Xeon(R) and 24 GB RAM and with a single 32-bit CPU, PANAM can process the phylogenetic analysis of 1000 eukaryotic OTUs of 400 bp in approximately 20 minutes, regardless of the affiliation method. The run time increased with the number of OTUs, regardless of the length. For example, for 400 bp, the run time ranged from 24 minutes for 5000 OTUs to 6 days and 14 hours for 1 M

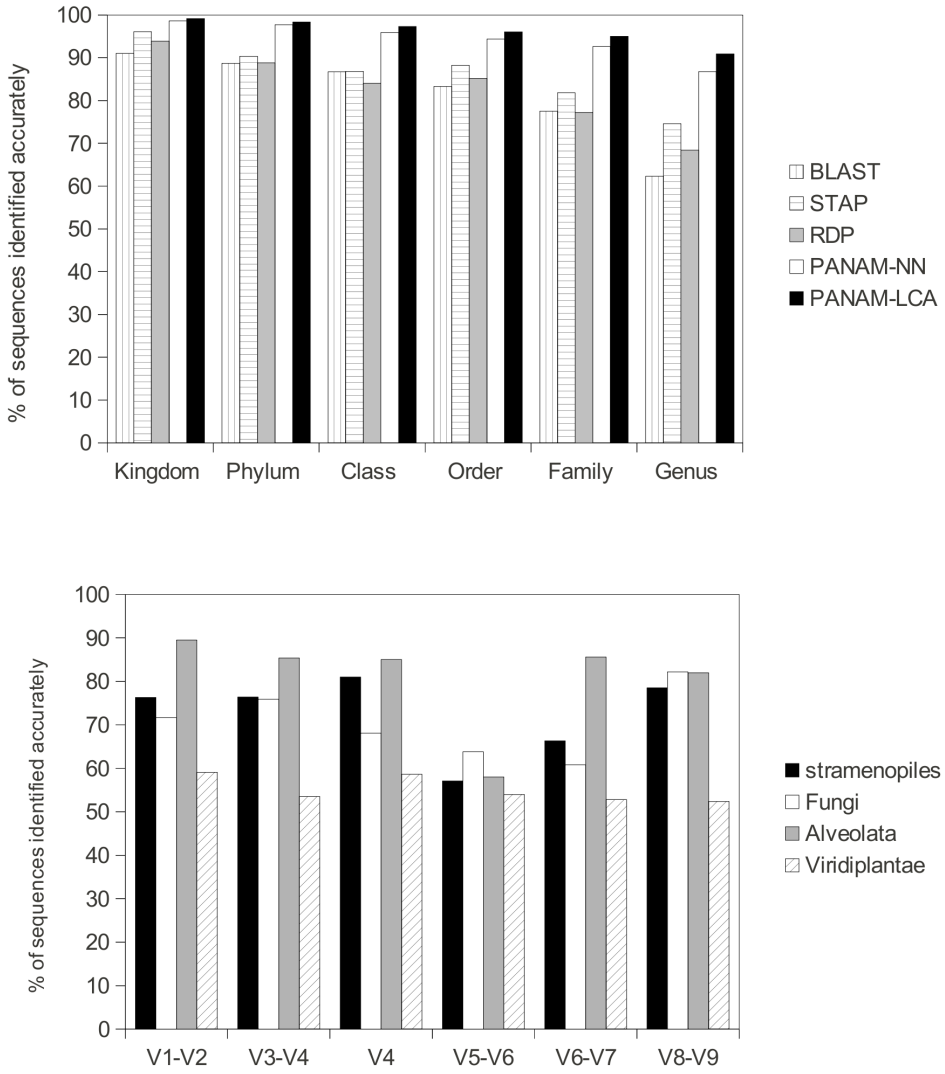


Figure 1. Accuracy of the phylogenetic affiliation of PANAM compared to different approaches and on different regions. 1.A. Accuracy of the phylogenetic affiliation of PANAM-LCA, PANAM-NN, STAP, BLAST and RDP Classifier. 1,000 near-full-length sequences were randomly picked from the reference database and removed from it for the simulations. For PANAM, simulations were repeated 5 times and the standard variation is less than 0.03. 1.B. Accuracy of the phylogenetic affiliation in relation with the variable region targeted. The specificity was tested with PANAM-LCA and a sequence length equal to 400 bp.
doi:10.1371/journal.pone.0058950.g001

eukaryotic OTUs. For near full-length sequences, PANAM was able to process 1 M sequences in 16 days (Figure S2).

Reliability of the phylogenetic affiliation in relation to the region targeted and the taxa of interest

The reliability of affiliations was compared for 400 bp reads spanning the 18 S rRNA gene for four taxonomic groups: Alveolata, Stramenopiles, Fungi and Viridiplantae at the genus level (Figure 1.B). Generally, the fragment affiliation depended on the taxonomic group and the region considered. According to previous results, the regions from V5 to V6 gave, on average, the weakest accuracy. Another general trend observed in this analysis was a poor taxonomic restitution for sequences belonging to Viridiplantae compared to other groups, between 52.4% and 59.1% regardless of the region targeted. The best specificity values for Stramenopiles, Alveolata and Fungi were obtained in different regions: V1–V2 (89.5%), V4 (81%), and V8–V9 (82.2%) respectively. The taxonomic affiliation for these three groups from the V8–V9 region was relatively similar, from 78.5% to 82.2%.

Tree-based analysis of pyrosequencing data from small lacustrine protists

In silico simulations have shown that primers NSF573 and NSR1147, used to target the V4 region of the 18S rRNA gene captured the greatest diversity (data not shown) and that the region amplified by these primers is suitable for taxonomic affiliation (Table 1). The reads were clustered at 95% similarity, and 6% of the OTUs (4% of reads) defined from this pyrosequencing run matched with Metazoa sequences and were not processed further. The diversity and richness indexes obtained for each environment are shown in Table 2. The lowest and highest richness indexes (Chao1) were found on Anterne Lake and Villerest Lake respectively, whereas the normalized indexes (based on 3759

Table 1. The specificity percentage values at the genus level for BLAST, RDP, STAP and PANAM (NN and LCA).

Starting position	Region	Length	BLAST	RDP	STAP	PANAM-NN	PANAM-LCA
NSF4	V1	200 bp	69.3	59.4	58.2	60.7	63.1
	V1–V2	400 bp	73.2	62.9	72	73.3	78.1
NSF370	V3	200 bp	61.7	54	54.2	55.9	50.2
	V3–V4	400 bp	70.9	67	70.8	70.2	73.3
NSF573	V4	200 bp	70.3	65.5	66.8	62.5	55.5
	V4	400 bp	72.3	67.8	69.9	74.6	76.8
NSF963	V5	200 bp	57.7	54.4	49.9	51.5	41.8
	V5–V6	400 bp	68.8	65.1	65.2	60.6	64.2
NSF1179	V6	200 bp	66.7	62.8	59.1	53.5	52.4
	V6–V7	400 bp	71.0	68.8	69.7	71.9	74.3
NSF1419	V8	200 bp	68.5	66.7	62.9	62.8	68.7
	V8–V9	400 bp	74.4	69.3	72.4	74	79.2

The specificity corresponds to the number of genus correctly affiliated among the detected ones, computed from forward primers for 200 bp and 400 bp amplicons. These values correspond to the mean computed from five samples of 1000 sequences (with the exception of V9 region computed with 300 sequences). The standard variation is less than 0.05.

doi:10.1371/journal.pone.0058950.t001

sequences) showed that Bourget Lake harboured the largest number of species (Table 2). This normalization also had an effect on the richness estimates in Godivelle Lake and Geneva Lake.

In the lakes studied, regarding level 2 and 3 from EMBL taxonomy (displayed in Table S3, a PANAM table output, including number of sequences, OTUs and diversity indexes), the major phylogenetic groups were Fungi, Alveolata and Stramenopiles representing 73.2% of OTUs and 78.6% of sequences (Figure 2). These mean values mask some disparities between lakes. Thus, Anterne Lake harboured mainly reads affiliated to Fungi (99.4% of total), whereas the main phylum in Geneva Lake was Alveolata (Figure 2; Table S3). Sequences belonging to the phylum Cryptophyta were the most abundant in Pavin Lake and Sep Lake.The results highlighted the presence of freshwater clades delineated in previous studies [6,37] such as Cryptophyta_2 to Cryptophyta_4, Rhizophydium or Cryptomycota (previously known as LKM11) among Fungi (Table S3). Sequences derived from Fungi, which were very abundant in sequence libraries from Anterne Lake and Aydat Lake, belonged to this last Cryptomycota clade (Table S3, Figure S3). These data demonstrate the presence of Chlorophyta and Haptophyta in all of the lakes studied, with the exception of Anterne Lake, which is characterised by an over-representation of Fungi and an absence of Haptophyta. This tree-based approach allows for the study of beta-diversity from phylogenies. The UNIFRAC metric showed that Bourget, Aydat and Anterne Lakes differed from other ecosystems regardless of the phylogenetic level (total Eukaryotes, Stramenopiles and Fungi) at which the analysis was performed (Figure 3).

In a comparison of the OTUs found in this study to those present in previous studies on the small protists, only 4.8% were previously detected in lakes. If only the dominant OTUs (>1% of reads) are taken into account, then the proportion of OTUs similar to specific lacustrine sequences increased to 19.7%. Moreover, new light is shed on putative clades of small protists. Specifically, these clades include the chlorophycean group of Mamiellophyceae, represented in Figure 4; Foraminifera (Rhizaria); Dictyochophyceae (Stramenopiles); and Euglenida (Euglenozoa). These clades were supported by high bootstrap values (> 0.8), included 23, 14, 17 and 23 OTUs respectively, and were found in at least three of the eight lakes. The novel clade within the Euglenozoa was composed only of OTUs present at less than 1% of reads.

Discussion

As the interplay between evolution and ecology receives more attention in ecosystem studies [38], there is greater interest in phylogenetic approaches for deciphering the mechanisms that govern the diversity and functioning of communities and ecosystems. However, the phylogenetic methods that are typically applied to Sanger-sequenced SSU rRNA are computationally expensive and cannot be readily used to handle NGS datasets; therefore, pyrosequencing reads are mainly analyzed by other approaches. The method described in this study is a response to the challenge of analyzing hundreds of thousands of SSU rRNA genes in a phylogenetic framework, inferring taxonomies from sister sequences and describing clades. This method has been implemented and tested for microorganisms with an emphasis on protists, which are not well served by bioinformatics tools dedicated to NGS data, although the early focus on bacterial and archaeal diversity has recently broadened to include eukaryotic microorganisms [39,40]; thus, the database provided in PANAM includes reference sequences from protists, Bacteria and Archaea and can be used for taxonomic assignment of all microorganisms.

Accuracy of affiliation methods for protist sequences

Our taxonomic affiliations were compared with BLAST, a tool commonly used for the identification of microorganisms especially microeukaryotes (e.g., [22]); RDP, which is currently used to classify bacterial and archaeal SSU rRNA sequences and fungal LSU rRNA sequences; and STAP implemented in WATERS [41]. This method, based on ClustalW alignments and PHYML phylogenies, is a standard method for taxonomic affiliations based on phylogenetic analyses. The RDP Classifier [42] is often considered to be restricted to bacterial and archaeal taxa [26] and therefore, is not used for eukaryotic classification of SSU rRNA genes after amplicon pyrosequencing. We used this tool for the first time for taxonomic affiliation of 18S rRNA gene amplicons generated with high-throughput pyrotag sequencing.

Table 2. Main characteristics of the lakes studied and richness and diversity indexes of small protists inferred from the pyrosequencing of amplicons .

Main characteristics			Richness and diversity						Richness and diversity normalized					
Lakes	Trophic status	Coordinates	Sequences	OTUs	Chao1	Shannon	ACE	Coverage	Sequences	OTUs	Chao1	Shannon	ACE	Coverage
Anterne	ultraoligotrophic	45°59′28″N, 6°47′54″E	17092	150	282.1	1.7	292.8	99.6	3759	51	93.0	0.5	121.3	99.3
Aydat	eutrophic	45°39′50″N, 2°59′04″E	8574	239	328.5	2.5	319.1	99.1	3759	176	235.1	2.59	237.6	98.5
Bourget	mesotrophic	45°43′55″N, 5°52′ 06″E	3759	294	442.6	4.0	478.6	96.7	3759	295	436.2	3.95	469.5	96.7
Geneva	mesotrophic	46°27′52″N, 6°33′31″E	10045	345	442.4	4.2	462.4	99.0	3759	158	199.0	3.70	203.5	98.9
Godivelle	ultraoligotrophic	45° 23′ 04″ N, 2° 55′ 25″ E	8742	234	317.8	3.8	313.2	99.2	3759	229	371.8	4.02	340.3	97.7
Pavin	oligomesotrophic	45°29′45″N, 2° 53′ 18″ E	11618	254	389.0	3.5	364.8	99.2	3759	157	287.7	3.39	244.7	98.4
Sep	oligomesotrophic	46° 02′ 51″ N, 3° 02′ 47″ E	7795	309	406.1	3.9	418.6	98.8	3759	232	329.5	3.79	322.4	98.0
Villerest	hypereutrophic	45° 59′ 36″ N, 4° 2′ 12″ E	8427	369	482.3	4.2	472.5	98.7	3759	277	399.5	4.14	373.9	97.4

doi:10.1371/journal.pone.0058950.t002

The affiliation of simulated amplicons were obtained by the RDP Classifier trained on the near full-length sequences of the reference database used in PANAM. Surprisingly, trimming the reference database to the primer region did not result in an improvement of classification for 18S rRNA gene sequences (data not shown), in contrast to the results of Werner et al. [43] on 16S rRNA gene sequences. As noted by these authors, a naïve Bayesian classification depends on the training set size. The weak performance on the truncated sequences could thus be explained by the limited number of 18S rRNA gene sequences in public databases compared with 16S rRNA gene sequences, particularly for the V9 region (see the discussion below).

The comparison of the tree-based method proposed with these tools in the context of taxonomic affiliation of 18S rRNA gene amplicons shows that regardless of the method that is used, taxonomic reliability depends on the sequence length and amplicon location on the SSU rRNA gene sequence. These results, which to our knowledge have not been examined for 18S rRNA gene sequences, are consistent with observations of 16S rRNA gene sequences from Bacteria and Archaea [44].

Our results mostly illustrate the impact of sequence length on phylogenetic methods, which appears to be the main limitation of this approach. According to Liu et al. [31], it is possible to use short fragments from the 16S rRNA gene to draw the same conclusions as with full-length sequences. However, by comparing different affiliation methods, they also noted that the short reads generated by pyrosequencing (i.e., 200 bp) were likely to be problematic for inferring phylogeny due to their small number of bases; similarity and probabilistic methods are therefore the most accurate. However, our analysis, similar to the one proposed by Jeraldo et al. [29] for 16S rRNA gene sequences, demonstrates that with the current average length achieved by the pyrosequencers (Titanium generation; $>$ 400 bp), phylogenetic methods are reliable and offer an advantage over other methods such as RDP. From 400 bp amplicons, the phylogenetic affiliation method implemented in PANAM outperforms the classical tools dedicated to NGS analysis at the genus level with the exception of amplicons sequences covering the V5–V6 region of the SSU rRNA gene. Phylogenetic methods are generally considered superior to other approaches for taxonomic affiliation [45] as they assess relatedness between a set of sequences. They are also considered to be difficult to automate as i) their reliability greatly depends on the quality of the alignments, which need to be validated by experts in the field, and ii) they use intensive, time-consuming methods for tree building.

In this study, we use the curated alignments sequences provided by SILVA, which is, at least for eukaryotic sequences, the only up-to-date curated database. All high-quality and near full-length aligned sequences suitable for in-depth phylogenetic analysis were selected. However, the guide-tree for eukaryotes provided by SILVA, in contrast to the other domains, represents only an approximate phylogeny. Tree-based approaches can implement other tools based on the tree-insertion methods like pplacer [46] as proposed by Bik et al. [28]. Similarly to STAP, this tool analyzes one sequence at a time. Thus, clades may be, at best, approximated from a frozen backbone tree, while the addition of distant taxa, as can be expected from environmental sequences, may require a re-evaluation of the phylogenetic tree [46]. In terms of processing time, we demonstrated that the tree-based method described here can process 1 M sequences in a reasonable (about three hours) time scale. For comparison, while pplacer processes 10,000 sequences in ~0.5 hour, PANAM can process 30,000 sequences in the same amount of time with the same computational resources. However, while a pyrosequencing run can

produce up to 1.2 M reads, the raw sequences first go through a quality control stage that eliminates poor quality reads and replicates. Additionally, in diversity studies, the raw sequences are first cleaned (i.e., quality trimmed) and clustered, and phylogenetic analyses are applied to the representatives of each OTU and not to all of the raw reads from a run. Consequently, in current studies of diversity, the effective number of sequences to be affiliated is on the order of tens of thousands, which can be processed in a few hours on a personal computer.

Accuracy of the protist affiliation in relation to the region targeted

The primers used for the taxonomic assignment of Bacteria traditionally span the regions V3, V6 and V9 of the SSU rRNA gene [12,47]. However, some studies [32,48] suggest that the V6 region is not optimal for taxonomic affiliation as it overestimates richness and the number of OTUs at different cut-offs [49]. In the microeukaryotic field, the regions V2–V3 [13], V3 [14,34], V4 [22,23,39] and V9 [21,22,24,25,39] were investigated with limited *in silico* analysis. Behnke et al. [39] partially addressed this concern because they compared the V4 and V9 regions for analyzing sequencing errors; V4 amplicons are likely more prone to an increased frequency of Roche 454 pyrosequencing homopolymer errors relative to the V9 region [22]. However, the inclusion of at least some part of the variable regions of the SSU RNA gene is necessary for the methods to retrieve sufficient signal for taxonomic affiliation. Liu et al. [32] stressed that tree-based methods are more sensitive to the 16S rRNA gene region targeted than are similarity-based methods because of different rates of evolution among regions [44], and/or the difference of homopolymer incidence and length between the regions [48]. The same conclusions can be drawn from our results from 18S rRNA gene amplicon sequences, because the accuracy of the phylogenetic affiliation for the region V5–V6 dropped for both phylogenetic methods used in this study (STAP and PANAM). Interestingly, the accuracy of the taxonomic affiliation of the main phyla varied with the region analyzed, but regardless of the variable region analyzed, simulated amplicons from Viridiplantae were always difficult to affiliate reliably at the genus level. Thus, the bias observed between variable regions [22] could be due to primers that may not anneal uniformly to all groups, but also to the bioinformatic process used for the taxonomic identification. In summary, with the exception of Viridiplantae, the V8–V9 region appears to be a

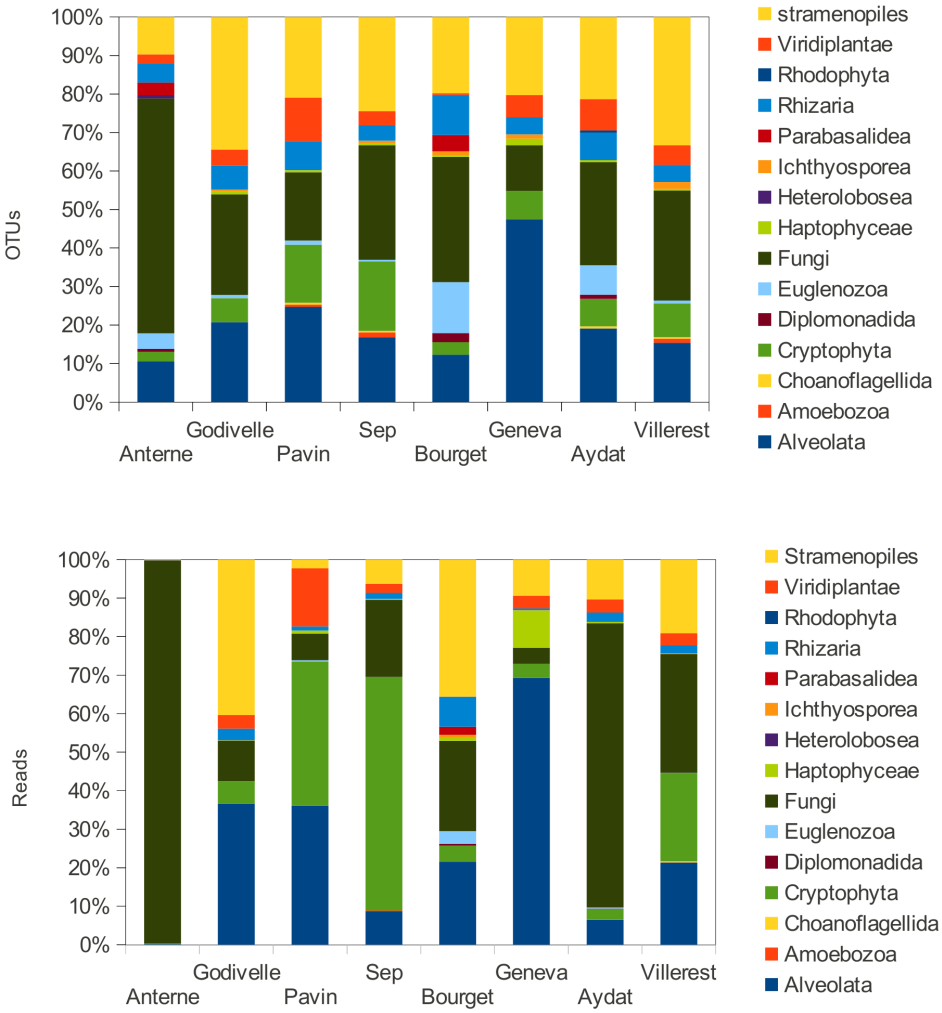


Figure 2. Proportions of the main phyla detected in the 8 lakes studied. The proportions are computed in term of OTUs (top) and reads (bottom) (see Table S3).
doi:10.1371/journal.pone.0058950.g002

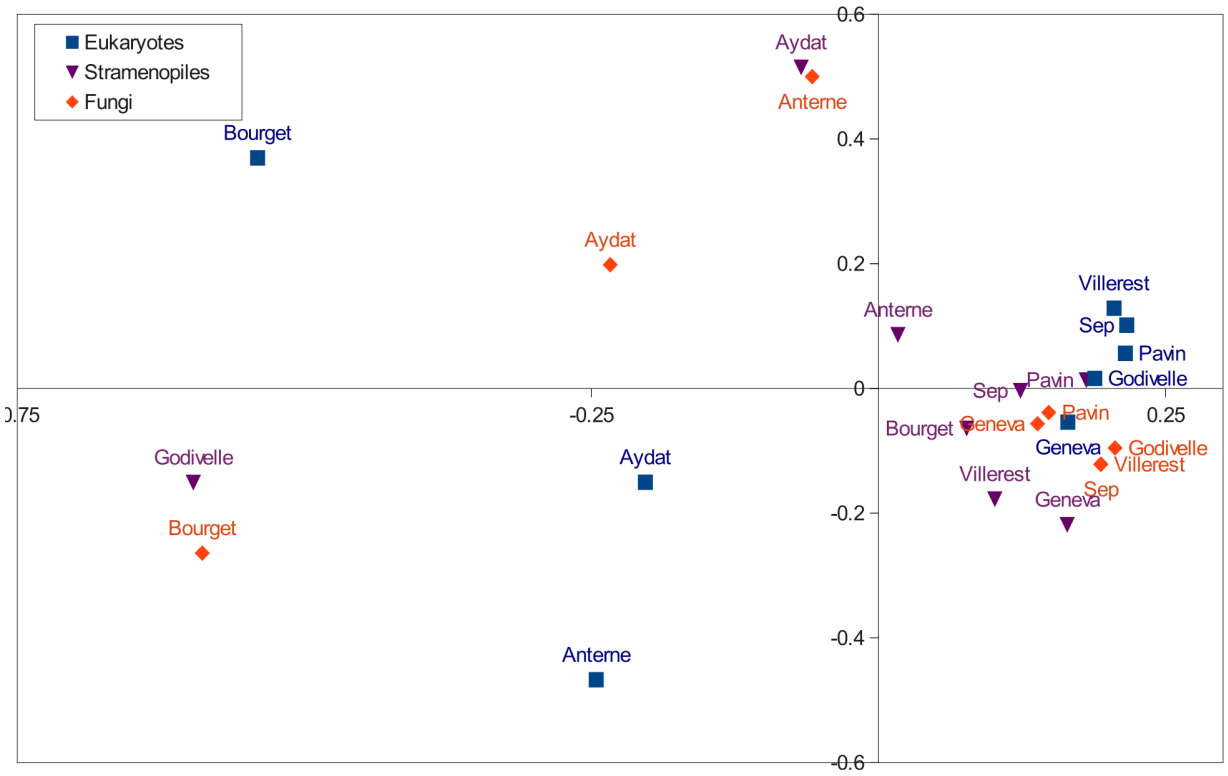


Figure 3. Principal coordinate analysis computed using a Unifrac distance metric from the phylogenies of the Stramenopiles, Fungi and the total eukaryotes. This analysis permit to differentiate environments according to their taxonomic composition. For example, Lake Godivelle seems to be different from the other lakes for the Stramenopiles, while it is similar for all eukaryotes.
doi:10.1371/journal.pone.0058950.g003

good candidate for the study of protist diversity because the reliability of the taxonomic affiliation did not differ according to the phyla considered (i.e., Stramenopiles, Fungi, Alveolata). However, sequence databases such as GenBank contain many fewer sequences that include the V9 region than other variable regions.

New insights into the small protist composition of the lacustrine ecosystem

In this analysis, our goal was not to explain the spatial pattern of the protist community composition (PCC) but to characterize the structure of these communities (richness, diversity and composition) by high-throughput SSU rRNA gene amplicon sequencing and sequence affiliation utilizing a tree-based method. We focused on the optimization of processing environmental data and on the

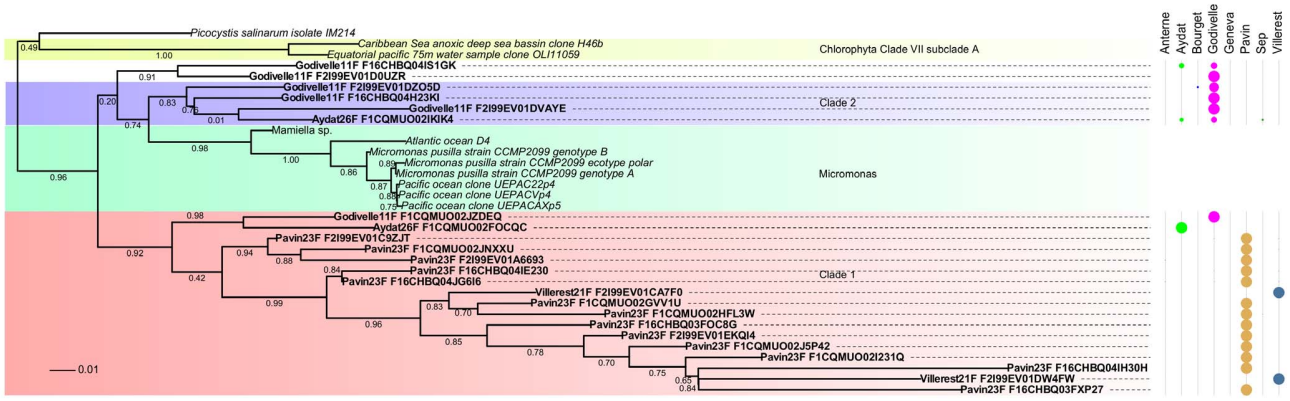


Figure 4. Main putative clades detected among Mamiellophyceae (Chlorophyceae) based on 18S SSU reads (425 bp ± 114). The OTUs affiliated to Chlorophyceae were generated at 95% similarity. A profile alignment was processed using HMMalign and the phylogeny was built by FASTTREE2 with 100 bootstraps. The distribution of the OTUs among different lakes shows a main presence of clade 1 in Lake Pavin while clade 2 is mainly present in Lake Godivelle.
doi:10.1371/journal.pone.0058950.g004

description of the general picture of protists diversity obtained for these lakes.

For an in-depth analysis of this PCC from lacustrine ecosystems, we introduced environmental sequences and taxonomies in the reference database to delineate specific clades as defined in previous publications (e.g., [6,37,50]). The introduction of "environmental reference" sequences reflecting the taxonomies of protists originating from specific environments can enhance the affiliation of poorly represented environmental sequences. Phylogenetic methods provide a clear edge in describing under-studied and complex communities. However, as with other methods, the precision of sequence mapping falls off when experimental sequences lie distant from reference SSU rRNA gene sequences [51]. This observation is particularly true for environmental sequences, for which the availability of close relatives and well-annotated sequences in reference databases is limited, as is the case for the V9 region. If the referenced trees do not include known relatives branching close to experimental reads, divergent lineages form long-branch taxa with no close reference sequences at relatively deep internal nodes. This phenomenon results in a less precise taxonomic affiliation of these sequences; however, clades of interest could still be drawn, as very similar sequences (i.e., sequences with low pairwise distance) are very well preserved among tree searches from *de novo* phylogenies [29].

Most eukaryotic species are defined on morphological differences, however, as the majority of existing microorganisms on Earth have not yet been cultured, their phenotypic traits can hardly be described. Thus, environmental microbial species are delineated according to a sequence similarity cut-off based on comparisons of SSU rRNA gene sequences to demarcate operational taxonomic units [52]. Although they do not technically represent species, OTUs composed of multiple sequences can be used to describe novel species, using the provisional designation of "Candidatus", when the SSU rRNA gene sequences are sufficiently different from those of recognized species [53]. In this study, after dataset cleaning and sorting, the reads left for the affiliation were clustered at a 95% identity threshold as proposed by Caron et al. [54] to delineate eukaryotic taxa. These authors defined this similarity threshold after studying the distribution of intra- and inter-specific variations of the 18S rRNA gene in protistan communities. However, as they pointed, this cut-off is a conservative estimator of species richness, and may mask considerable physiological diversity in some OTUs. In other studies, taxon clustering is performed at sequence similarity from 90% to 100% [23]. As the error rate of many NGS platforms in any case is $\sim$1% it is recommanded to cluster at a lower threshold than 99%. Some authors chose a similarity of 97% because this value is commonly used to define OTUs in Bacteria (e.g., [22]). However, this value has been defined for delineating a species from the full-length 16S rRNA gene. Thus, from *in silico* analysis of 16S rRNA genes, Kim et al. [33] showed that the clustering threshold must be chosen according to the variable region amplified and the domain studied (i.e., Archaea or Bacteria). A less conservative cut-off could overestimate the richness and diversity because in some phyla, such as diatoms, the level of intragenomic polymorphism in the SSU rRNA gene can reach 2% [55]. Finally, in a previous study, Mangot et al. [56] defined a threshold of 95 % by adding an internal standard (a clonal sequence derived from a copy of the 18S rRNA gene in Blastocystis subtype 4 genome) before amplifying and sequencing the DNA samples. Indeed, all the amplicons derived from this sequence clustered in one OTU at this cut-off.

Our tree-based treatment applied to NGS sequences demonstrated that few OTUs have been previously described by the traditional cloning-sequencing (CS) method. As these OTUs represent taxa present in relatively low abundance in many environments, little information is available about them. These novel OTUs were contained in a broad range of higher level taxa, including i) well-established clades such as Cryptomycota, ii) in phyla rarely detected by cultivation-independent sequencing (e.g., Ichthyosporea) and iii) in novel clades previously undescribed in lacustrine ecosystems, such as Foraminifera.

Thus, according to this study, the OTUs representing the most abundant sequences were found among Fungi, Alveolata, Stramenopiles, Cryptophyta and Rhizaria. More precisely, the phylogenetic affiliation allows to delineate three of the four previously defined freshwater Cryptophyta clades [6]. Within the Fungi, numerous OTUs were associated with Cryptomycota [57] or Chytridiomycota, which include both parasitic and saprotrophic organisms [58]. The presence of Chlorophyta and Haptophyta was confirmed in most of the lake environments sampled in this study. By the CS method used for describing PCC, Chlorophyta and Haptophyta were often absent [59,60] or found at a very low proportion [6,37], whereas these phyla represented a significant proportion of PCC when counting methods such as FISH were used [61]. Such a bias has also been highlighted in marine environments since epifluorescence microscopy reveals a dominance of phototrophic or mixotrophic cells over heterotrophic cells [62]. Another example of phyla rarely described yet detected here is the Ichthyosporea phylum, which was found only in hyper-eutrophic conditions [63]. Finally, some clades supported by high bootstrap values in our phylogenies, e.g., Mamiellales or Foraminifera, seem original because they have not been detected by CS with 'universal' eukaryotic primers. To our knowledge this is the first time that a clade closely associated to Mamiellales, as defined by Marin and Melkonian [64], has been detected in lakes. Present but scarce in our pyrosequencing data, these microalgae constitute the dominant photosynthetic group among the picoplankton 18S rRNA gene sequences in marine surveys ($\sim$ 1/3 of the sequences), especially in coastal waters, and have been shown to account for 45% of the picoeukaryotic community, as targeted by TSA-FISH in these waters [65,66]. The freshwater counterpart of this group, the Monomastigales, is rarely recovered from environmental samples and likely requires new molecular approaches that will specifically target photosynthetic organisms in the environment [64]. Freshwater Foraminifera, a group of granuloreticulosan protists largely neglected until now have already been detected by using specific primers in one study of freshwater ecosystems [67]. Thus, a NGS sequencing analysis with a moderate depth ($\sim$ 10,000 cleaned read per sample for Eukaryota) allows for the detection of the main phylogenetic phyla but also rarely detected phyla or phyla only detected by specific primers which act similar to massively parallel sequencing by focusing on one clade. Among the biases commonly assigned to CS, other than the variability in the cell lysis efficiency, the rRNA gene copy number, which range from 1 to 12,000 [68] is certainly the most important and may result in an over-representation of heterotrophic organisms notably of the alveolate taxa [34]. However, even if these differences in copy number distort the interpretation in number of reads and OTUs for both the CS and NGS methods, the massively parallel sequencing can at least increase detection of rare lineages or organisms with low gene copy numbers thanks to the increased depth of sequencing. We can hypothesize that this copy number could be more homogeneous at a specific lower taxonomic level (for example Alveolata), and the various indexes were therefore computed for each phylum instead of considering the whole protistan community (Table S3).

Conclusion

These results show that phylogenetic methods provide a clear edge in describing under-studied and complex communities, allowing the taxonomic affiliation of experimental sequences within an evolutionary framework; the study of relatedness among both environmental and reference sequences; and the evaluation of proximity of experimental sequences (''binning''). Thus, the tree-based method presented in this work, applied to the whole spectrum of microorganisms diversity (i.e., Eukaryota, Bacteria and Archaea), makes it possible to seek typical clades, allowing for the discovery of new putative lineages that are rarely or never recovered by classical sequencing approaches and the investigation of specific features within ecosystems considering sampling depths and periods. This feature cannot be inferred with a similarity search, a naïve Bayesian classification (RDP) or tree-based methods that process one sequence at a time.

Materials and Methods

The data originating from simulations and pyrosequencing were processed by a pipeline, referred to as PANAM (Phylogenetic Analysis of Next-generation AMplicons) that is based on publicly available programs. In addition to the phylogenetic analysis, this pipeline allows for the complete analysis of a full pyrosequencing run, including raw data processing, sequence clustering into OTUs and generating phylogenies for the taxonomic affiliation. The description of the procedure is detailed in the following sections (*''Processing of raw pyrosequencing reads and OTU picking''; ''Phylogenetic affiliation''; ''Richness and diversity indexes''*). It is written in Perl and can be run on Linux. The package comprises a reference sequence database, a taxonomy file and reference profile alignments and can be obtained from http://code.google.com/p/panam-phylogenetic-annotation/.

Processing of raw pyrosequencing reads and OTU picking

The pyrosequencing reads can be cleaned according to different methods commonly used in the field of molecular microbial ecology. Pyrosequencing errors can therefore be reduced by removing the primers (e.g., [69]), defining a minimal score and length of the reads (e.g., [14]) or removing reads with unidentified bases (Ns).

Short sequences and sequences with low-quality scores are removed using PANGEA scripts [16] and only sequences with a primer match percentage above a defined threshold are selected using Fuznuc [70]. Alternatively, other quality filtering methods can be implemented; the platform does not depend upon the filtering approach described above. When several samples are analyzed, the checked sequences are split into different files depending on their bar code or tag. Then, generated files are clustered using USEARCH [71] at a user-defined threshold, and representative sequences from OTUs are selected for the phylogenetic assignment.

Phylogenetic affiliation

For the phylogenetic affiliation, a dedicated database of reference sequences, verified taxonomy and alignments was built using sequences extracted from the SSURef 108 database of the SILVA project [72]. For this purpose, all the sequences (16S and 18S rRNA genes) with more than 1,200 bp, quality score $> 75\%$, and a pintail value > 50 were extracted. The sequence quality score defined by SILVA is a combination of the percentages of ambiguities, homopolymers longer than 4 bases and possible vector contaminations, and the pintail value corresponds to the probability that the rRNA sequence is chimeric. The complete database, after filtering according to the criteria above, contains 164,353 sequences (Archaea: 11,092; Bacteria: 131,428; and Eukaryota: 21,833) together with their taxonomy. To speed up the phylogenetic processing, the 3 domains were split into 37 phyletic groups of unicellular organisms corresponding to the first monophyletic clade after domains, as annotated in the guide-tree of SILVA (ARB format), and clustered at 97% identity.

Each profile corresponds to the first rank beneath that of domain. As the taxonomy of Bacteria and Archaea follow standardized taxonomic paths, the monophyletic profiles of these two domains correspond to phylum, the first level occurring after the domain. For Eukaryota domain, the taxonomy does not necessarily fit this organization, and the position of the taxon in the taxonomic hierarchy does not imply rank as it is the case with Bacteria and Archaea. Therefore, for the eukaryotic profiles, we opted for the rank position (the first one after the eukaryotic domain) and the monophyly, regardless to the taxonomic level.

For each of the 37 phyletic groups, an outgroup containing one sequence from each other group belonging to the same domain plus 2 external sequences were added to the alignment to root the phyletic tree to be produced and to specify the relatedness of early diverging sequences from the root of the group. To broaden the targeted diversity, the user can add specific environmental sequences to the database and the profiles.

Using this dedicated database, the phylogenetic affiliation is carried out following the different stages described in the Figure 5.

1- First, OTUs are compared against the reference database described above with USEARCH [71]. As this first step does not intend to provide an exact affiliation, but rather to give a first approximation to perform a rapid and accurate phylogenetic analysis, the query sequences are sorted according to the taxonomy of their best hits, whatever their similarity score. Several files are generated, each containing the reads and their 5 best hits, assigned to one of the 37 specific phyletic groups.

2- After reads have been assigned to phyletic groups, they are aligned to the reference sequences of the corresponding profile alignment for that group using hmmalign from the HMMER package [73]. Synthetic files, which include the reference sequences and the aligned experimental reads, are generated.

3- Using FASTTREE [74], a bootstrapped phylogenetic tree (100 iterations) is built for each phyletic profile, including OTUs associated with their 5 best hits and the reference sequences. The trees are then parsed to generate files containing the taxonomy of the inserted sequences and files reporting the clades that could be identified from reads forming monophyletic groups. Two methods for taxonomy assessment are implemented: lowest common ancestor (LCA) and nearest neighbor (NN). In this last method, for each query sequence, all the nodes containing the sequence are scanned from the most recent to the deepest. The closest neighbor is defined as the first referenced sequence starting from the lowest node. The query sequence will acquire the complete taxonomy of its nearest neighbor. For LCA [32] each node holds only the common taxonomy between all of its descendants and thus may be incomplete. Each query sequence will inherit the taxonomy of its lowest node. The final taxonomy assignment is based on the phylogeny. The relatedness between all sequences (both experimental and referenced) are re-evaluated, and the similarity based assignments proposed on stage 1 are therefore revised to provide a more phylogeny-driven affiliation. Regarding the clades, their definition differs according to authors (e.g., [75,76]), although in general, a new clade is declared when the cluster contains environmental sequences from at least 3 different sources and is

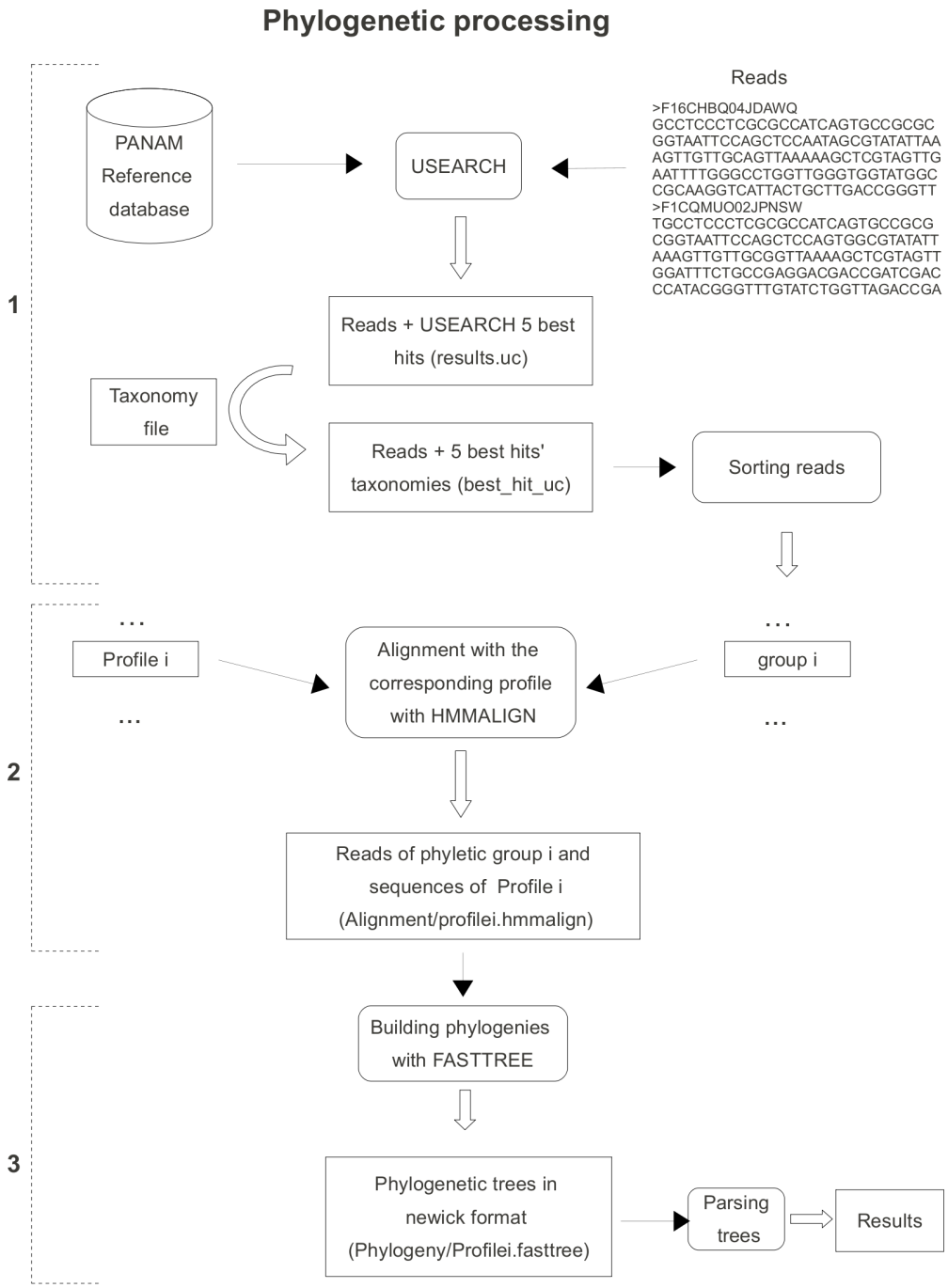


Figure 5. Flow chart describing the phylogenetic affiliation. A primary classification, sorts and splits reads into groups according to the taxonomy of their best USEARCH hit (1). Next, a file containing aligned reads and sequences from the corresponding group is generated by processing a profile alignment by HMMER. This file is used by FASTTREE to build a phylogenetic tree (2), which is then parsed to assign a taxonomy to each read and to report putative clades (3).
doi:10.1371/journal.pone.0058950.g005

supported by bootstrap values generally higher than 70%. The files generated describe monophyletic clusters with all the information required for experts in the field to define a putative environmental clade: a bootstrap value, a list of all the experimental sequences affiliated to it and the nearest reference neighbour together with its taxonomy. The implementation of PANAM (files generated) is extensively described in the documentation associated with the pipeline.

Richness and diversity indexes

After the cleaning step, richness (Chao1 and ACE), diversity (Shannon) indexes, and coverage are computed for each sample [77]. Subsequently, sequence library sizes are equalized to avoid

biases associated with different sampling depths (e.g., [78]). Briefly, the same number of sequences (i.e., the number of sequences in the smallest sample) are randomly sampled from each library, and diversity indexes are calculated for these equalized datasets. After phylogenetic affiliation, Chao1 and the Shannon diversity indexes are computed for levels 2 and 3 from the EMBL classification (e.g., Stramenopiles and Bacillarophyta).

Analysis of sequencing data obtained from simulations

PANAM was first tested on near full-length sequences with known taxonomy using 5 sets of 1000 sequences randomly picked from the reference database and removed from it for evaluations to be re-affiliated. The reliability of PANAM taxonomic affiliations was evaluated for specificity defined as the proportion of ranks correctly affiliated among the detected ones. A pyrosequencing simulation was also performed with pseudo-reads being generated by clipping the 5 × 1000 full-length sequences datasets from 6 universal forward primers for Eukaryotes [79] (Table S2). Clipped sequences were extended 200 and 400 bp from the forward primer positions defined on the *Saccharomyces cerevisiae* sequence (V01335), thus covering regions with different variability along the 18S rRNA gene. As emphasized, this pipeline allows taxonomic affiliations within an evolutionary context: its performance was thus primarily compared with that of STAP (Small Subunit rRNA Taxonomy and Alignment Pipeline) [51], the phylogenetic affiliation method used in WATERS (Workflow for the Alignment, Taxonomy, and Ecology of Ribosomal Sequences) [41], but was also compared with non-phylogenetic methods, including BLAST and the RDP Classifier implemented in MOTHUR [19] trained on the near full-length and trimmed sequences of the reference database.

The computational load of the phylogenetic analyses using PANAM was also tested with increasingly large datasets to evaluate processing time on a personal computer and to detect any scaling issues.

Analysis of sequencing data obtained from environmental studies

The PANAM tree-based method was run on environmental sequences, namely i) a set of environmental sequences originating from published studies on the diversity of protists and belonging to described environmental lacustrine clades of Perkinsozoa and Cercozoa [6,7] and ii) from an environmental survey of the lacustrine protist diversity performed in eight freshwater ecosystems.

For this purpose, eight lakes or reservoirs, described in Table 2 (Lakes Anterne, Aydat, Bourget, Godivelle, Geneva, and Pavin, and Reservoirs Sep and Villerest), were sampled once during their thermal stratification (from May to August according to the lake). Water samples from the epilimnion (1 to 5 m) were collected with a Van Dorn bottle at a permanent station (the deepest zone of the lake). Water samples (from 100 to 120 ml) were successively filtered through 5 µm-pore-size and 0.2 µm-pore-size polycarbonate filters (Millipore), and the membranes were stored at-80°C until nucleic acid extraction. All samples were extracted following the protocol described previously by Lefranc et al. [37].

The V4-V5 variable region of eukaryotic 18S rDNA was amplified with primers Ek-NSF573 and Ek-NSR1147 (Table S2). To discriminate each sample, a 5 bp multiplex tag was coupled with the Roche 454 pyrosequencing adaptor A. The amplification mix (30 µl) contained 30 ng of genomic DNA, 200 µM of deoxynucleoside triphosphate (Bioline, London, UK), 2 mM MgCl2 (Bioline), 10 pmol of each primer, 1.5 U of *Taq* DNA polymerase (Bioline) and the PCR buffer. The cycling conditions were an initial denaturation at 94°C for 10 min followed by 30 cycles of 94°C for 1 min, 57°C for 1 min, 72°C for 1 min and 30 s and a final 10-min extension at 72°C. Finally, the amplicons of all of the samples were pooled at equimolar concentrations and pyrosequenced using a Roche 454 GS-FLX system (Titanium Chemistry) by GATC (Konstanz, Germany). The reads, alignments and trees have been deposited in Dryad (http://datadryad.org). The reads used in this study were selected from a full run, separated into bins according to the tags, analyzed by PANAM, using trimming criteria of quality score > 22 and sequence length > 200 bases and clustering into OTUs with a 95% similarity threshold. UNIFRAC metrics [30] and a principal coordinate analysis were used to compare the small protist community between the lakes based on phylogenetic information obtained by PANAM using the packages Picante and ade4 implemented in the R software [80].

To broaden the covered diversity, more specifically regarding the environmental and pyrosequencing datasets processed in this study, and to build phylogenies with more similar sequences for the studied environment, 173 sequences from eukaryotic clades specific to lacustrine ecosystems, defined in previous works (e.g., [6,37]), were introduced in the eukaryotic reference database and the corresponding groups.

Supporting Information

Figure S1 The Cercozoa (A) and Perkinsea (B) phylogenies generated by PANAM after inserting environmental sequences. Inserted environmental sequences are in color (sequences with no accession number have been deposited in GenBank).
(PDF)

Figure S2 Processing time of PANAM-LCA depending on the number and length of reads.
(PDF)

Figure S3 The Cryptomycota phylogeny displaying the representative OTUs detected in the lakes. A representative OTU can be picked from a particular ecosystem but can be present in all ecosystems sampled as the OTU named Anterne08F F1CQMUO02ICISV.
(PDF)

Table S1 Comparison of the different approaches of taxonomic assignment.
(PDF)

Table S2 The primers names and sequences used in the simulations and pyrosequencing.
(PDF)

Table S3 Main taxonomic groups with richness and diversity indexes in the different lakes studied.
(PDF)

Author Contributions

Conceived and designed the experiments: NT GB DD. Performed the experiments: NT JFM ID GB DD. Analyzed the data: NT GB JFM ID DD. Contributed reagents/materials/analysis tools: NT ID GB DD. Wrote the paper: ID GB DD.

References

1. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3: REVIEWS0003.
2. Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. Science 296: 1061–1063. doi:10.1126/science.1070710.
3. Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, et al. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420: 806–810. doi:10.1038/nature01240.
4. DeLong EF (1992) Archaea in coastal marine environments. Proc Natl Acad Sci USA 89: 5685–5689.
5. López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature 409: 603–607. doi:10.1038/35054537.
6. Lepère C, Domaizon I, Debroas D (2008) Composition of freshwater small eukaryotes community: unexpected importance of potential parasites. Applied and Environmental Microbiology.
7. Mangot J-F, Debroas D, Domaizon I (2011) Perkinsozoa, a well-known marine protozoan flagellate parasite group, newly identified in lacustrine systems: a review. Hydrobiologia 659: 37–48. doi:10.1007/s10750-010-0268-x.
8. Galand PE, Casamayor EO, Kirchman DL, Lovejoy C (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. Proc Natl Acad Sci USA 106: 22427–22432. doi:10.1073/pnas.0908284106.
9. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL (2011) Activity of Abundant and Rare Bacteria in a Coastal Ocean. Proc Natl Acad Sci USA 108: 12776–12781. doi:10.1073/pnas.1101405108.
10. Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1: 283–290. doi:10.1038/ismej.2007.53.
11. Brown MV, Philip GK, Bunge JA, Smith MC, Bissett A, et al. (2009) Microbial community structure in the North Pacific ocean. The ISME Journal 3: 1374–1386.
12. Andersson AF, Lindberg M, Jakobsson H, Bäckhed F, Nyrén P, et al. (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS ONE 3: e2836. doi:10.1371/journal.pone.0002836.
13. Monchy S, Sanciu G, Jobard M, Rasconi S, Gerphagnon M, et al. (2011) Exploring and quantifying fungal diversity in freshwater lake ecosystems using rDNA cloning/sequencing and SSU tag pyrosequencing. Environmental Microbiology 13: 1433–1453. doi:10.1111/j.1462-2920.2011.02444.x.
14. Nolte V, Pandey RV, Jost S, Medinger R, Ottenwälder B, et al. (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Mol Ecol 19: 2908–2915. doi:10.1111/j.1365-294X.2010.04669.x.
15. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37: D141–D145. doi:10.1093/nar/gkn879.
16. Giongo A, Crabb DB, Davis-Richardson AG, Chauliac D, Mobberley JM, et al. (2010) PANGEA: pipeline for analysis of next generation amplicons. ISME J 4: 852–861. doi:10.1038/ismej.2010.16.
17. Pandey RV, Nolte V, Schlötterer C (2010) CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. BMC Res Notes 3: 3. doi:10.1186/1756-0500-3-3.
18. Mori H, Maruyama F, Kurokawa K (2010) VITCOMIC: visualization tool for taxonomic compositions of microbial communities based on 16S rRNA gene sequences. BMC Bioinformatics 11: 332. doi:10.1186/1471-2105-11-332.
19. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75: 7537–7541. doi:10.1128/AEM.01541-09.
20. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335–336. doi:10.1038/nmeth.f.303.
21. Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora M, et al. (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. BMC Biology 7: 72. doi:10.1186/1741-7007-7-72.
22. Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, et al. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Molecular Ecology 19: 21–31. doi:10.1111/j.1365-294X.2009.04480.x.
23. Cheung MK, Au CH, Chu KH, Kwan HS, Wong CK (2010) Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. The ISME Journal 4: 1053–1059. doi:10.1038/ismej.2010.26.
24. Pawlowski J, Christen R, Lecroq B, Bachar D, Shahbazkia HR, et al. (2011) Eukaryotic Richness in the Abyss: Insights from Pyrotag Sequencing. PLoS ONE 6: e18169. doi:10.1371/journal.pone.0018169.
25. Edgcomb V, Orsi W, Bunge J, Jeon S, Christen R, et al. (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. ISME J 5: 1344–1356.
26. Bik HM, Sung W, De Ley P, Baldwin JG, Sharma J, et al. (2012) Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. Molecular Ecology 21: 1048–1059. doi:10.1111/j.1365-294X.2011.05297.x.
27. von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, et al. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315: 1126–1130. doi:10.1126/science.1133420.
28. Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, et al. (2012) Sequencing our way towards understanding global eukaryotic biodiversity. Trends in Ecology & Evolution 27: 233–243. doi:10.1016/j.tree.2011.11.010.
29. Jeraldo P, Chia N, Goldenfeld N (2011) On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys. Environ Microbiol 13: 3000–3009. doi:10.1111/j.1462-2920.2011.02577.x.
30. Lozupone CA, Knight R (2008) Species divergence and the measurement of microbial diversity. FEMS Microbiol Rev 32: 557–578. doi:10.1111/j.1574-6976.2008.00111.x.
31. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R (2007) Short Pyrosequencing Reads Suffice for Accurate Microbial Community Analysis. Nucl Acids Res 35: e120. doi:10.1093/nar/gkm541.
32. Liu Z, DeSantis TZ, Andersen GL, Knight R (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res 36: e120. doi:10.1093/nar/gkn491.
33. Kim M, Morrison M, Yu Z (2011) Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. Journal of Microbiological Methods 84: 81–87. doi:10.1016/j.mimet.2010.10.020.
34. Medinger R, Nolte V, Pandey RV, Jost S, Ottenwälder B, et al. (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Molecular Ecology 19: 32–40. doi:10.1111/j.1365-294X.2009.04478.x.
35. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
36. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
37. Lefranc M, Thénot A, Lepère C, Debroas D (2005) Genetic diversity of small eukaryotes in lakes differing by their trophic status. Appl Environ Microbiol 71: 5935–5942. doi:10.1128/AEM.71.10.5935-5942.2005.
38. Schoener TW (2011) The newest synthesis: understanding the interplay of evolutionary and ecological dynamics. Science 331: 426–429. doi:10.1126/science.1193954.
39. Behnke A, Engel M, Christen R, Nebel M, Klein RR, et al. (2010) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. Environmental Microbiology 13: 340–349. doi:10.1111/j.1462-2920.2010.02332.x.
40. Dawson S, Hagen K (2009) Mapping the protistan "rare biosphere." Journal of Biology 8: 105. doi:10.1186/jbiol201.
41. Hartman A, Riddle S, McPhillips T, Ludascher B, Eisen J (2010) Introducing W.A.T.E.R.S.: a Workflow for the Alignment, Taxonomy, and Ecology of Ribosomal Sequences. BMC Bioinformatics 11: 317. doi:10.1186/1471-2105-11-317.
42. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261–5267. doi:10.1128/AEM.00062-07.
43. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, et al. (2012) Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. ISME J 6: 94–103.
44. Schloss PD (2010) The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. PLoS Comput Biol 6: e1000844. doi:10.1371/journal.pcbi.1000844.
45. Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26: 1641–1650. doi:10.1093/molbev/msp077.
46. Matsen F, Kodner R, Armbrust EV (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11: 538. doi:10.1186/1471-2105-11-538.
47. Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Welch DM, et al. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet 4: e1000255. doi:10.1371/journal.pgen.1000255.
48. Bowen De León K, Ramsay B, Fields M (2012) Quality-Score Refinement of SSU rRNA Gene Pyrosequencing Differs Across Gene Region for Environmental Samples. Microbial Ecology: 1–10. doi:10.1007/s00248-012-0043-9.
49. Youssef N., Sheik CS., Krumholz LR., Najar FZ., Roe BA., Elshahed MS. (2009) Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys. Appl. Environ. Microbiol. August 2009 vol. 75 no. 16 5227-5236
50. Richards TA, Vepritskiy AA, Gouliamova DE, Nierzwicki-Bauer SA (2005) The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. Environ Microbiol 7: 1413–1425. doi:10.1111/j.1462-2920.2005.00828.x.

51. Wu D, Hartman A, Ward N, Eisen JA (2008) An Automated Phylogenetic Tree-Based Small Subunit rRNA Taxonomy and Alignment Pipeline (STAP). PLoS ONE 3: e2566. doi:10.1371/journal.pone.0002566.
52. Ward DM, Cohan FM, Bhaya D, Heidelberg JF, Kühl M, et al. (2007) Genomics, environmental genomics and the issue of microbial species. Heredity 100: 207–219. doi:10.1038/sj.hdy.6801011.
53. Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nature Reviews Microbiology 6: 431–440. doi:10.1038/nrmicro1872.
54. Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, et al. (2009) Defining DNA-Based Operational Taxonomic Units for Microbial-Eukaryote Ecology. Appl Environ Microbiol 75: 5797–5808. doi:10.1128/AEM.00298-09.
55. Alverson AJ, Kolnick L (2005) Intragenomic nucleotide polymorphism among small subunit (18s) rDNA paralogs in the diatom genus skeletonema (bacillariophyta)1. Journal of Phycology 41: 1248–1257. doi:10.1111/j.1529-8817.2005.00136.x.
56. Mangot J-F, Domaizon I, Taib N, Marouni N, Duffaud E, et al (2013) Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. Environmental Microbiology. doi: 10.1111/1462-2920.12065
57. Jones MDM, Forn I, Gadelha C, Egan MJ, Bass D, et al. (2011) Discovery of novel intermediate forms redefines the fungal tree of life. Nature 474: 200–203. doi:10.1038/nature09984.
58. Rasconi S, Jobard M, SimeNgando T (2011) REVIEW Parasitic fungi of phytoplankton: ecological roles and implications for microbial food webs. Aquat Microb Ecol 62: 123–137. doi:10.3354/ame01448.
59. Lepère C, Boucher D, Jardillier L, Domaizon I, Debroas D (2006) Succession and regulation factors of small eukaryote community composition in a lacustrine ecosystem (Lake Pavin). Applied and environmental microbiology 72: 2971.
60. Tarbe A, Stenuite S, Balagu V, Sinyinza D, Descy J, et al. (2011) Molecular characterisation of the small-eukaryote community in a tropical Great Lake (Lake Tanganyika, East Africa). Aquat Microb Ecol 62: 177–190. doi:10.3354/ame01465.
61. Lepère C, Masquelier S, Mangot J-F, Debroas D, Domaizon I (2010) Vertical structure of small eukaryotes in three lakes that differ by their trophic status: a quantitative approach. ISME J 4: 1509–1519.
62. Not F, del Campo J, Balagué V, de Vargas C, Massana R (2009) New Insights into the Diversity of Marine Picoeukaryotes. PLoS ONE 4: e7143. doi:10.1371/journal.pone.0007143.
63. Lepère C, Domaizon I, Debroas D (2007) Community composition of lacustrine small eukaryotes in hyper-eutrophic conditions in relation to top-down and bottom-up factors. FEMS Microbiol Ecol 61: 483–495. doi:10.1111/j.1574-6941.2007.00359.x.
64. Marin B, Melkonian M (2010) Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear- and plastid-encoded rRNA operons. Protist 161: 304–336. doi:10.1016/j.protis.2009.10.002.
65. Vaulot D, Eikrem W, Viprey M, Moreau H (2008) The diversity of small eukaryotic phytoplankton (≤3 μm) in marine ecosystems. FEMS Microbiology Reviews 32: 795–820. doi:10.1111/j.1574-6976.2008.00121.x.
66. Not F, Latasa M, Marie D, Cariou T, Vaulot D, et al. (2004) A single species Micromonas pusilla (Prasinophyceae) dominates the eukaryotic picoplankton in the western English Channel. Appl Environ Microbiol 70: 4064-4072.
67. Holzmann M, Habura A, Giles H, Bowser SS, Pawlowski J (2003) Freshwater foraminiferans revealed by analysis of environmental DNA samples. J Eukaryot Microbiol 50: 135–139.
68. Zhu F, Massana R, Not F, Marie D, Vaulot D (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS Microbiology Ecology 52: 79-92.
69. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc Natl Acad Sci U S A 103: 12115–12120. doi:10.1073/pnas.0605127103.
70. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16: 276–277. doi:10.1016/S0168-9525(00)02024-2.
71. Edgar RC (2010) Search and Clustering Orders of Magnitude Faster Than BLAST. Bioinformatics 26: 2460–2461. doi:10.1093/bioinformatics/btq461.
72. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188–7196. doi:10.1093/nar/gkm864.
73. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763.
74. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5: e9490. doi:10.1371/journal.pone.0009490.
75. Zwart G, Crump BC, Agterveld MPK, Hagen F, Han S (2002) Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. Aquat Microb Ecol 28: 141–155. doi:10.3354/ame028141.
76. Groisillier A, Massana R, Valentin K, Vaulot D, Guillou L (2006) Genetic diversity and habitats of two enigmatic marine alveolate lineages. Aquat Microb Ecol 42: 277–291. doi:10.3354/ame042277.
77. Hill TCJ, Walsh KA, Harris JA, Moffett BF (2003) Using ecological diversity measures with bacterial communities. FEMS Microbiol Ecol 43: 1–11. doi:10.1111/j.1574-6941.2003.tb01040.x.
78. Gihring TM, Green SJ, Schadt CW (2012) Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. Environmental Microbiology 14: 285–290. doi:10.1111/j.1462-2920.2011.02550.x.
79. Van de Peer Y, Robbrecht E, de Hoog S, Caers A, De Rijk P, et al. (1999) Database on the structure of small subunit ribosomal RNA. Nucleic Acids Res 27: 179–183.80. R Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Primer	Sequence
NSF4	CTGGTTGATYCTGCCAGT
NSF370	AGGGYTCGAYYCCGGAGA
NSF573	CGCGGTAATTCCAGCTCCA
NSF963	TTRATCAAGAACGAAAGT
NSF1179	AATTTGACTCAACACGGG
NSF1419	ATAACAGGTCTGTGATGCCC
NSR1147	CCGTCAATTYYTTTRAGTTT

Table S1: The primers names and sequences used in the simulations and pyrosequencing

		Anterne					Aydat					Bourget					Godivelle				
		#seq	#OTUs	Schao1	Shannon	Coverage	#seq	#OTUs	Schao1	Shannon	Coverage	#seq	#OTUs	Schao1	Shannon	Coverage	#seq	#OTUs	Schao1	Shannon	Coverage
Alveolata		**32**	**13**	**28.00**	**2.39**	**81.25**	**498**	**35**	**50.00**	**2.72**	**97.99**	**564**	**26**	**39.00**	**0.81**	**97.52**	**2450**	**50**	**69.12**	**1.92**	**99.27**
	Apicomplexa	0	0				38	2				42	3				12	5			
	Ciliophora	29	12				245	20				14	9				2278	33			
	Dinophyceae	3	1				183	11				27	11				149	10			
	Perkinsea	0	0				32	2				481	3				11	2			
Amoebozoa		**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	Centramoebida	0	0				0	0				0	0				0	0			
	Tubulinea	0	0				0	0				0	0				0	0			
	unclassified Amoebozoa	0	0				0	0				0	0				0	0			
Choanoflagellida		**0**	**0**				**2**	**1**	**1.00**	**0.00**	**100.00**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	unclassified Choanoflagellida	0	0				2	1				0	0				0	0			
Cryptophyta		**9**	**3**	**3.00**	**1.00**	**100.00**	**210**	**13**	**23.00**	**1.23**	**97.62**	**108**	**7**	**13.00**	**0.94**	**96.30**	**389**	**15**	**17.00**	**1.51**	**98.97**
	Cryptomonadaceae	2	1				11	2				11	1				207	4			
	Cryptophyta nucleomorph	0	0				4	2				1	1				21	4			
	Cryptophyta_2	5	1				5	1				0	0				10	3			
	Cryptophyta_3	2	1				42	4				95	4				54	3			
	Cryptophyta_4	0	0				148	4				1	1				97	1			
Diplomonadida		**1**	**1**	**1.00**	**0.00**	**0.00**	**2**	**2**	**3.00**	**0.69**	**0.00**	**14**	**5**	**5.00**	**1.51**	**92.86**	**0**	**0**	**0**	**0**	**0**
	Hexamitidae	1	1				2	2				14	5				0	0			
Euglenozoa		**6**	**5**	**8.00**	**1.56**	**33.33**	**24**	**14**	**18.20**	**2.54**	**70.83**	**85**	**28**	**35.86**	**2.97**	**87.06**	**2**	**2**	**3.00**	**0.69**	**0.00**
	Euglenida	6	5				24	14				85	28				2	2			
Fungi		**16946**	**75**	**109.36**	**1.65**	**99.83**	**5604**	**49**	**68.12**	**0.70**	**99.68**	**613**	**69**	**115.87**	**2.06**	**93.80**	**704**	**63**	**84.08**	**2.34**	**96.73**
	Chytridiomycota	0	0				11	2				3	2				6	4			
	Dikarya	10	8				27	14				80	32				80	22			
	Glomeromycota	0	0				0	0				0	0				2	1			
	LKM11	16893	56				5411	20				20	2				27	5			
	Microsporidia	0	0				0	0				0	0				2	1			
	Rhizophydium clade	1	1				1	1				0	0				19	4			
	Rozella clade	0	0				32	1				0	0				8	3			
	Zygomycota_1 et rel.	0	0				12	1				0	0				3	1			
	unclassified Fungi	42	10				110	10				510	33				557	22			
Haptophyceae		**0**	**0**	**0**	**0**	**0**	**23**	**1**	**1.00**	**0.00**	**100.00**	**24**	**1**	**1.00**	**0.00**	**100.00**	**3**	**2**	**2.00**	**0.64**	**66.67**
	Pavlovales	0	0				0	0				0	0				1	1			
	unclassified Haptophyceae	0	0				23	1				24	1				2	1			
Heterolobosea		**2**	**1**	**1.00**	**0.00**	**100.00**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	Schizopyrenida	2	1				0	0				0	0				0	0			
Ichthyosporea		**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**18**	**2**	**2.00**	**0.45**	**100.00**	**2**	**1**	**1.00**	**0.00**	**100.00**
	unclassified Ichthyosporea	0	0				0	0				18	2				2	1			
Parabasalidea		**8**	**4**	**7.00**	**1.07**	**62.50**	**0**	**0**	**0**	**0**	**0**	**53**	**9**	**11.00**	**1.35**	**92.45**	**0**	**0**	**0**	**0**	**0**
	Trichomonada	8	4				0	0				53	9				0	0			
Rhizaria		**9**	**6**	**9.00**	**1.68**	**55.56**	**183**	**13**	**13.33**	**1.62**	**98.91**	**204**	**22**	**31.00**	**1.69**	**95.10**	**202**	**15**	**25.50**	**1.62**	**96.53**
	Cercozoa	4	2				173	10				168	6				201	14			
	Foraminifera	5	4				10	3				36	16				1	1			
Rhodophyta		**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1.00**	**0.00**	**0.00**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	Florideophyceae	0	0				1	1				0	0				0	0			
Viridiplantae		**3**	**3**	**6.00**	**1.10**	**0.00**	**259**	**15**	**17.00**	**1.80**	**98.46**	**1**	**1**	**1.00**	**0.00**	**0.00**	**233**	**10**	**11.00**	**0.87**	**98.71**
	Chlorophyta	3	3				259	15				1	1				233	10			
stramenopiles		**21**	**12**	**30.00**	**2.17**	**57.14**	**785**	**39**	**50.00**	**2.42**	**98.60**	**930**	**42**	**53.00**	**2.84**	**98.71**	**2700**	**83**	**86.93**	**3.36**	**99.59**
	Bacillariophyta	7	1				14	2				0	0				14	5			
	Bicosoecida	1	1				1	1				187	16				364	21			
	Chrysophyceae	9	6				221	16				489	12				1908	37			
	Dictyochophyceae	1	1				32	6				17	3				280	9			
	Eustigmatophyceae	2	2				270	2				0	0				0	0			
	Labyrinthulida	0	0				15	3				32	3				55	4			
	Oomycetes	0	0				91	3				65	3				3	1			
	unclassified stramenopiles	1	1				141	6				140	5				76	6			

Tab. S2 : Main taxonomic groups with richness and diversity indices in the different lakes studied

Tab

Geneva					Pavin					Sep					Villerest				
#seq	#OTUs	Schao1	Shannon	Coverage	#seq	#OTUs	Schao1	Shannon	Coverage	#seq	#OTUs	Schao1	Shannon	Coverage	#seq	#OTUs	Schao1	Shannon	Coverage
4838	**84**	**99.83**	**2.79**	**99.59**	**2476**	**46**	**57.00**	**2.32**	**99.52**	**569**	**42**	**49.50**	**2.85**	**98.24**	**1036**	**42**	**53.00**	**2.52**	**98.94**
11	2				3	2				109	7				293	9			
4276	59				1565	32				395	23				330	21			
548	21				903	10				18	8				73	7			
3	2				5	2				47	4				340	5			
0	**0**	**0**	**0**	**0**	**3**	**1**	**1.00**	**0.00**	**100.00**	**10**	**3**	**3.00**	**0.94**	**90.00**	**3**	**3**	**6.00**	**1.10**	**0.00**
0	0				0	0				0	0				1	1			
0	0				0	0				0	0				1	1			
0	0				3	1				10	3				1	1			
0	**0**	**0**	**0**	**0**	**1**	**1**	**1.00**	**0.00**	**0.00**	**1**	**1**	**1.00**	**0.00**	**0.00**	**13**	**1**	**1.00**	**0.00**	**100.00**
0	0				1	1				1	1				13	1			
255	**13**	**14.50**	**1.53**	**98.82**	**2561**	**28**	**46.00**	**1.75**	**99.65**	**3948**	**45**	**56.25**	**2.10**	**99.75**	**1111**	**24**	**31.00**	**1.76**	**99.37**
13	3				63	4				1202	15				177	6			
3	1				345	6				67	7				13	4			
167	4				50	3				2552	18				670	6			
5	3				1692	13				112	3				250	7			
67	2				411	2				15	2				1	1			
0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
0	0				0	0				0	0				0	0			
0	**0**	**0**	**0**	**0**	**22**	**2**	**2.00**	**0.30**	**100.00**	**1**	**1**	**1.00**	**0.00**	**0.00**	**4**	**2**	**2.00**	**0.56**	**75.00**
0	0				22	2				1	1				4	2			
292	**21**	**36.00**	**1.27**	**96.58**	**474**	**33**	**40.50**	**2.21**	**97.89**	**1303**	**74**	**87.57**	**3.26**	**98.47**	**1496**	**78**	**86.08**	**3.41**	**99.00**
0	0				0	0				27	3				26	3			
12	7				323	21				75	14				65	14			
0	0				1	1				0	0				0	0			
15	2				14	2				136	9				13	2			
0	0				0	0				0	0				0	0			
0	0				0	0				244	11				550	23			
8	2				0	0				53	3				7	1			
0	0				0	0				3	1				49	1			
257	10				136	9				765	33				786	34			
678	**3**	**3.00**	**0.25**	**100.00**	**56**	**1**	**1.00**	**0.00**	**100.00**	**13**	**1**	**1.00**	**0.00**	**100.00**	**1**	**1**	**1.00**	**0.00**	**0.00**
0	0				0	0				0	0				0	0			
678	3				56	1				13	1				1	1			
0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
0	0				0	0				0	0				0	0			
3	**2**	**2.00**	**0.64**	**66.67**	**0**	**0**	**0**	**0**	**0**	**2**	**2**	**3.00**	**0.69**	**0.00**	**8**	**5**	**11.00**	**1.39**	**50.00**
3	2				0	0				2	2				8	5			
0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
0	0				0	0				0	0				0	0			
32	**8**	**9.50**	**1.70**	**90.62**	**72**	**14**	**28.00**	**1.90**	**88.89**	**103**	**10**	**10.75**	**1.12**	**97.09**	**101**	**12**	**15.00**	**1.98**	**97.03**
23	4				67	9				30	7				101	12			
9	4				5	5				73	3				0	0			
0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
0	0				0	0				0	0				0	0			
231	**10**	**10.00**	**2.01**	**99.57**	**1033**	**21**	**22.50**	**1.04**	**99.61**	**148**	**9**	**10.50**	**1.60**	**97.97**	**153**	**14**	**24.50**	**1.77**	**95.42**
231	10				1033	21				148	9				153	14			
655	**36**	**49.00**	**2.18**	**98.02**	**156**	**39**	**67.50**	**3.04**	**87.82**	**412**	**61**	**74.57**	**3.21**	**95.15**	**928**	**91**	**107.87**	**3.47**	**97.52**
356	6				2	2				6	4				190	9			
3	2				53	16				30	12				279	45			
250	17				71	13				159	17				297	17			
8	2				1	1				145	13				4	3			
1	1				1	1				0	0				0	0			
26	3				3	2				19	3				34	3			
4	2				23	3				30	5				89	4			
7	3				2	1				23	7				35	10			

. S2 (continued) : Main taxonomic groups with richness and diversity indices in the different lakes studied

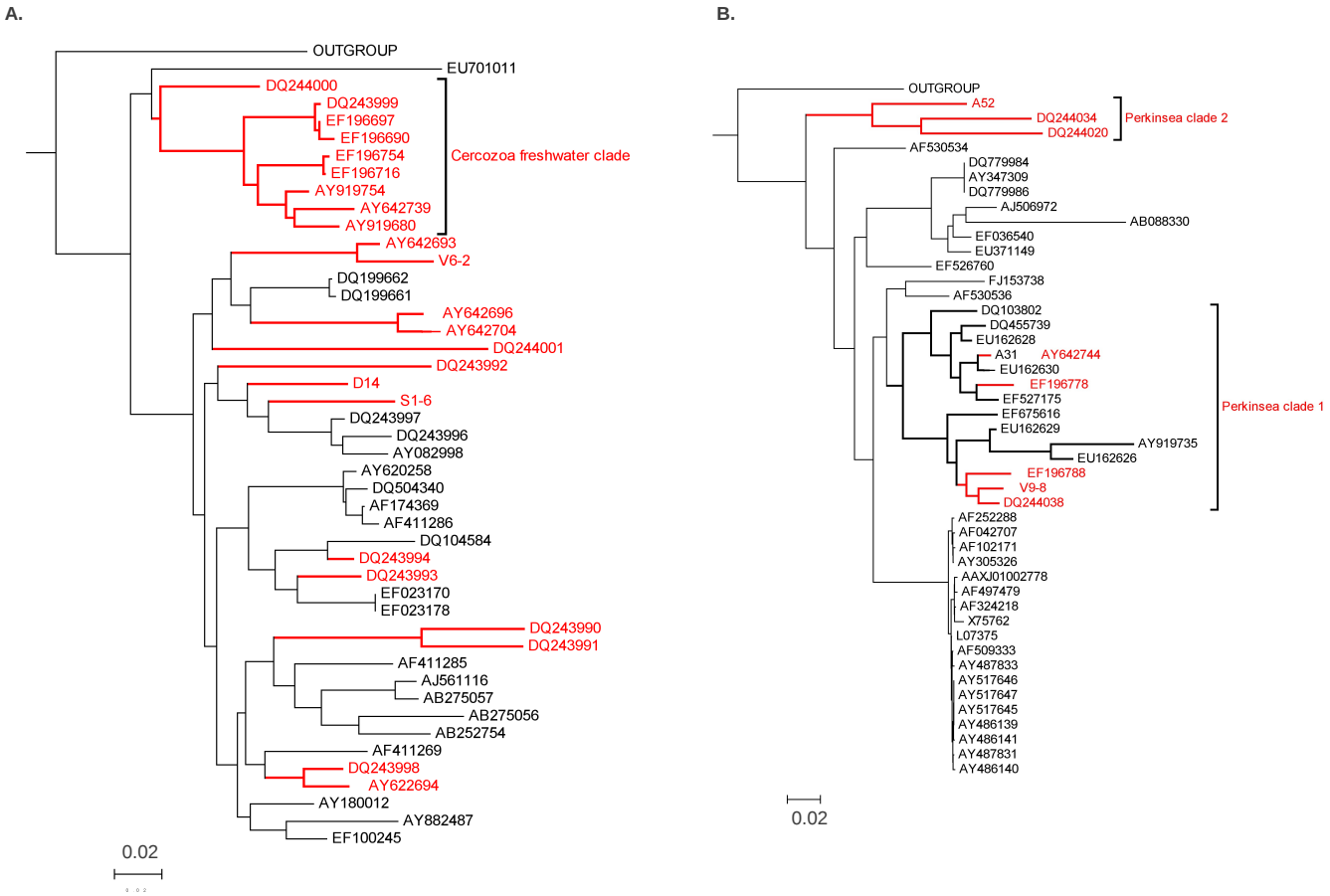


Fig. S1: The *Cercozoa* (A) and *Perkinsea* (B) phylogenies generated by PANAM after inserting environmental sequences. Inserted environmental sequences are in color (sequences with no accession number have been deposited in GenBank).

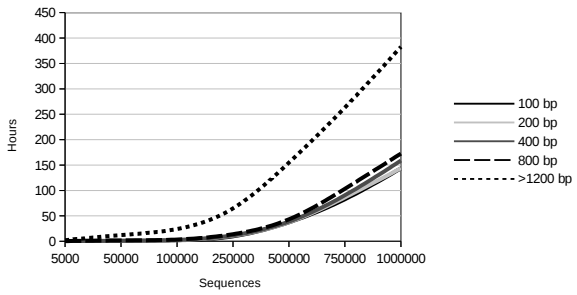


Fig. S2: Processing time of PANAM-LCA depending on the number and length of reads.

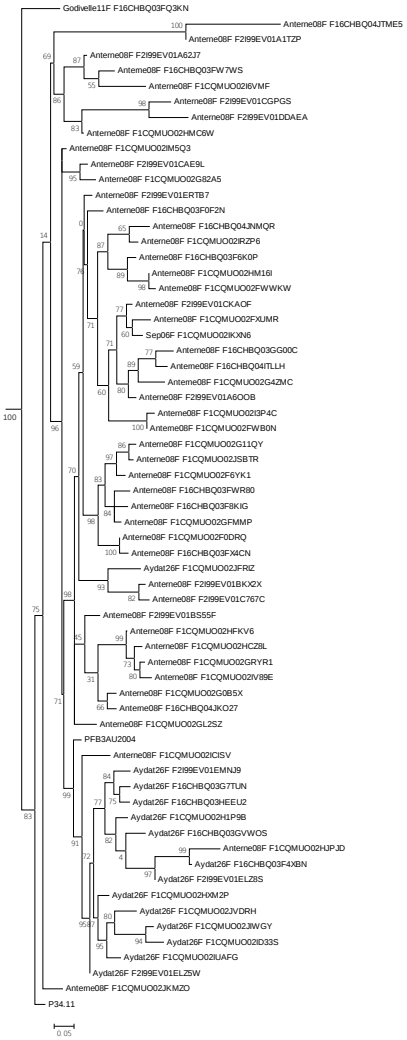


Fig. S3: The Cryptomycota phylogeny displaying the representative OTUs detected in the lakes. A representative OTU can be picked from a particular ecosystem but can be present in all ecosystems sampled as the OTU named Anterne08F F1CQMUO02ICISV.

Article 2

ePANAM : a web server for depicting the microbial diversity from high throughput amplicons

(Article en préparation)

Application note

ePANAM: a web server for depicting the microbial diversity from high-throughput sequencing amplicons

Najwa Taib[1,5*], Gisèle Bronner[1,5], Jean-Christophe Charvy[1,5], Simon Roux[1,5], Mylène Hugoni[1,5], Antoine Mahul[2], Engelbert Mephu-Nguifo[3,6], Vincent Breton[4,7], Didier Debroas[1,5]

[1]Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes : Génome et Environnement", BP 10448, F-63000 CLERMONT-FERRAND FRANCE, [2]Clermont Université, Université Blaise Pascal,CRRI, [3]Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 CLERMONT-FERRAND FRANCE, [4]Clermont Université, Université Blaise Pascal, Laboratoire de Physique Corpusculaire, BP 10448, F-63000 CLERMONT-FERRAND FRANCE, [5]CNRS, UMR 6023, LMGE, F-63171 AUBIERE FRANCE, [6]CNRS, UMR 6158, LIMOS, F-63171 AUBIERE FRANCE, [7]CNRS, UMR 6533, LPC, F-63171 AUBIERE FRANCE

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: High-throughput data production has revolutionized microbial ecology, allowing for deep insights into environmental diversity. However, the massive increases in the data generation capacity require automated and computationally intensive approaches. Herein, we describe ePANAM, a web server dedicated to the phylogenetic analysis of next generation amplicons that provides the user with a suite of tools deployed on a computing cluster to process bacterial, archaeal and microeukaryotic SSU rDNA reads. ePANAM processes the raw outputs from pyrosequencers to depict the structure of the microbial community (alpha- and beta- diversity) and taxonomic composition of single or multiplexed samples through a phylogeny-based method. The ePANAM outputs consist of an Operational Taxonomic Units list, diversity indices, taxonomy tables, phylogenies and a list of putative new clades detected in the sample(s).

Availability: ePANAM is freely available online at http://panam-meb.univ-bpclermont.fr

Contact:

Supplementary information:

1 INTRODUCTION

The characterization of microbial community structure via SSU rRNA gene profiling (Woese and Fo, 1977) has been greatly advanced in recent years by the introduction of amplicon pyrosequencing. In addition, the possibility of barcoding provides the opportunity to massively screen multiple samples from different environments in a single run, leading to a better representation of sample diversity at a lower cost. This progress in method development has enabled the study of alpha, beta, and gamma diversity (e.g., Youssef et *al.*, 2010; King et *al.*, 2012). Moreover, it has provided a new window into the composition of microbial communities (Neufeld et *al.*, 2005; Sogin et *al.*, 2006) and sparked interest in the members of rare biospheres (e.g., Sogin et *al.*, 2006).

A multitude of command line tools can be used for the basic processing steps of these data, such as quality checking, OTU selection and taxonomy assignment (e.g., MOTHUR (Schloss et *al.*, 2009); RDP (Cole et *al.*, 2009); QIIME (Caporaso et *al.*, 2010)). However, freely available tools that offer intuitive user interfaces and hide the details of installation and computing costs are still scarce. RDP and PyroTagger (Kunin and Hugenholtz, 2010) partially fill in this blank by offering a web-interface pipeline for pyrotag analysis. However, although phylogenetic affiliation has been shown to be more accurate for the taxonomic assignment of pyrosequencing reads (Taib et *al.*, 2013), these tools assign taxonomy by either a similarity search or probabilistic approach, with the phylogenies being restricted to beta-diversity estimates.

We present ePANAM, a web server that performs a large-scale phylogeny-based analysis of the 16S and 18S rRNA genes, with no need for specialized informatics expertise. This server is based on a validated pipeline for pyrosequencing outputs (Taib et *al.*, 2013), which had been used to depict archaeal and picoeukaryotic community structures and dynamics (Hugoni et *al.*, 2013; Lepère et *al.*, 2013; Mangot et *al*., 2013). After a cleaning step to remove low quality sequences, ePANAM sorts sequences by their barcode identifiers. Diversity and richness estimators are then calculated within each sample, as well as rank abundance and rarefaction curves. The OTUs selected from each sample are processed within a phylogenetic framework. These phylogenies enable the assessment of taxonomy and delineation of monophyletic groups to highlight clades of interest. The ePANAM graphical representation of the results makes it easy to obtain an overview of the sample diversity. The computations are performed on a computing cluster, allowing for multiple parallel runs.

2 METHODS

As the input, ePANAM takes files containing sequences (FASTA), quality scores (qual) and barcodes. The web interface allows users to choose the

*To whom correspondence should be addressed.

settings for the analysis: quality filters, OTU selection cutoff, sequence primers, and normalization threshold for *Bacteria*, *Archaea* or Eukarya (Protists and Fungi). The results are displayed under three tabs (alpha-diversity, beta-diversity and taxonomy composition). The parameters set by the user are summarized on a meta data page.

2.1 Processing raw sequences

ePANAM first pre-processes the pyrosequencing dataset to remove short sequences and trim those sequences that contain bases with low quality scores using the PANGEA script (Giongo et *al.*, 2010). Next, the trimmed sequences are examined for primer matches and the presence of ambiguous bases (Ns). Once the low quality sequences are discarded according to the user instructions, the remaining sequences are sorted according to the barcodes, followed by removal of the barcodes and adaptors. The sequences of each sample are either pooled or analyzed independently to compute the OTUs using UCLUST (Edgar, 2010) at the user-defined threshold and with the optimal settings.

2.2 Generating phylogenies and assigning taxonomy

First, the OTUs are compared with the PANAM databases using USEARCH and sorted according to the higher taxonomy level (e.g., phylum) of their best hits. Files containing the OTUs together with their five best hits are then generated. Each file is aligned to the reference alignment of the taxonomic group it has been associated with via an HMM profile using HMMER (Eddy, 1998); the alignments are then used to generate phylogenetic trees with 100 bootstraps using FastTree2 (Price et *al.*, 2010). Finally, the generated trees are rooted and parsed. The nearest neighbor (NN) taxonomy (e.g., STAP (Wu et *al*., 2008)) and the taxonomy of the lowest common ancestor (LCA) of the OTU's representative sequence (e.g., Huson et *al.*, 2007) are inferred to the query sequences according to their position in the phylogeny. The phylogenies are also used to describe monophyletic groups of environmental sequences that may form clades.

2.3 Describing alpha-diversity

For each sample, the OTUs computed at the processing step are used to calculate taxonomy-independent diversity and richness indices (Chao1, Shannon, ACE and Coverage index) to generate rarefaction curves that can be used to determine whether the sampling effort covers the entire community and to generate rank abundance curves. To avoid biases associated with different sampling depths, these values are calculated for both the whole data set in a sample and for a normalized size library of sequences that are randomly selected from the sample (Gihring et *al.*, 2011). Moreover, diversity and richness indices (Chao1 and Shannon) are calculated for each taxonomic group in a sample at two levels (e.g., phylum and class), and pie charts displaying the taxonomic richness (OTUs) and abundance (reads) are generated.

2.4 Describing beta-diversity

To compare different environments, ePANAM computes the OTUs that are common among the samples. A table is generated with the occurrence of the OTUs in each sample and their taxonomies (NN and LCA) for both the whole and normalized data sets. For taxonomy-independent comparisons, these contingency tables are then used to plot correspondence analyses (CA) and visualize the distribution of samples in terms of their OTU compositions. A taxonomy-dependent comparison is realized using heatmaps to discriminate samples according to their taxonomic composition and abundance. All the graphs are generated using R software (R Development Core Team 2012).

3 IMPLEMENTATION

The ePANAM computations are distributed on a cluster, allowing multiple parallel runs. The processing schedule is driven by a workflow manager that relies on an XML workflow description file, job states files and several basic scripts. The workflow manager is built on twelve steps structured around two utility operations. After raw sequence processing in five bricks (quality features, OTU definition, profile trimming, sequence sorting and data parallelization drafting), the first technical operation generates as many child threads as the parallelization drafting has defined parcels of data. Afterwards, each bunch of sequences is aligned against the reference bases, and the phylogenetic trees are constructed. Then, phylogeny parsing and clade definitions are performed. The second technical operation aims to gather and merge the thread results before building the summary files about the diversity indices or taxonomic distributions. At the end, a script packs the computing results into an archive file, and an e-mail is sent to notify the user that the data processing is completed.

ACKNOWLEDGEMENTS

Funding:This work was supported by the french Conseil Régional d'Auvergne.
Conflict of interest: none declared.

REFERENCES

Caporaso,J.G., et al. *(2010) QIIME allows analysis of high-throughput community sequencing data.* Nat Methods, *7: 335–336.*

Cole,J.R., et al. *(2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis.* Nucleic Acids Res, *37, D141–D145.*

Core Team (2012) R: A language and environment for statisticalcomputing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Eddy,S.R. (1998) Profile hidden Markov models. Bioinformatics, *14, 755–763.*

Edgar,R.C. (2010) Search and Clustering Orders of Magnitude Faster Than BLAST. Bioinformatics, *26, 2460–2461.*

Gihring,T.M., et al. (2012) Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. Environ Microbiol, *14(2):285-90.*

Giongo,A., et al. *(2010) PANGEA: pipeline for analysis of next generation amplicons.* ISME *J, 4, 852–861.*

Hugoni,M., et al. *(2013) Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters.* Proc Natl Acad Sci U S A., *110(15):6004-9. doi: 10.1073/pnas.1216863110*

Huson,D.H., et al. *(2007) MEGAN analysis of metagenomic data.* Genome Res, *17(3), 377-86.*

King,G.M., et al. *(2012) Analysis of composition and structure of coastal to mesopelagic bacterioplankton communities in the northern gulf of Mexico.* Front Microbiol, *3, 438.*

Kunin,V. and Hugenholtz,P. (2010) PyroTagger: a fast, accurate pipeline for analysis for analysis of rRNA amplicon pyrosequencing data. The Open J *1: 1–8*

Lepère,C., et al. (2013) Geographic distance and ecosystem size determine the distribution of smallest protists in lacustrine ecosystems. FEMS Microbio Ecol. *Doi: 10.1111/1574-6941.*

Mangot,J.F., et al. *(2013) Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes.* Environmental Microbiology. *Doi: 10.1111/1462-2920.12065*

Neufeld,J.D. and Mohn,W.W. (2005) Unexpectedly high bacterial diversity in arctic tundra relative to boreal forest soils, revealed by serial analysis of ribosomal sequence tags. Appl Environ Microbiol, *71(10), 5710-5718.*

Price,M.N., et al. *(2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.* PLoS *ONE 5: e9490. doi:10.1371/journal.pone.0009490.*

Schloss,P.D., et al. *(2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.* Appl Environ Microbiol, *75, 7537–7541.*

Sogin,M.L. et al. *(2006) Microbial diversity in the deep sea and the underexplored "rare biosphere".* Proc. Natl. Acad. Sci., *103(32), 12115-12120.*

Taib,N., et al. *(2013) Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity.* PLoS One, *8(3):e58950. doi: 10.1371/journal.pone.0058950*

Woese,C.R. and Fox,G.E. (1977) Phylogenetic structure of the prokaryotic domain: The primary *kingdoms*. Proc. Natl. Acad. Sci., *74(11), 5088-5090.*

Wu,D., et al. *(2008) An Automated Phylogenetic Tree-Based Small Subunit rRNA Taxonomy and Alignment Pipeline (STAP).* PLoS *ONE, 3: e2566. doi:10.1371/journal.pone.0002566.*

Youssef,N.H., et al. *(2010) Fine-scale bacterial beta diversity within a complex ecosystem (Zodletone Spring, OK, USA): the role of the rare biosphere.* PloS ONE, *8, e12414.*

Chapitre 3

Applications

3.1 Introduction

Dans ce chapitre, nous avons appliqué notre méthode de traitement à des données de NGS provenant de deux contextes différents, en collaboration avec deux doctorants : M. Hugoni et J-F. Mangot. Ces études d'écologie microbienne concernaient deux modèles différents, les *Archaea* et les eucaryotes, les deux fractions provenant d'écosystèmes aquatiques. Ces études avaient pour point commun de décrypter la dynamique de la biosphère rare. La première (article 3) concernait plus précisément la caractérisation des communautés d'*Archaea* en milieu marin. Nous avons ainsi analysé un jeu de séquences issues des données du pyroséquençage de l'ADNr 16S et de l'ARNr 16S et provenant de 40 dates d'échantillonnage, qui nous a permis d'étudier les remaniements de la biosphère rare en différenciant les micro-organismes actifs et inactifs. La seconde étude (article 4) concernait la dynamique à court terme des picoeucaryotes (organismes dont la taille est inférieure à 5 μm) dans un écosystème lacustre. Les études réalisées sur le moyen et long terme ont montré que les assemblages eucaryotiques se restructurent rapidement pour suivre une distribution rang-abondance en log-normal, avec quelques taxa abondants (Vigil et al., 2009, Nolte et al., 2010). Selon Caron and Countway (2009), ces remaniements continus au sein de la communauté des protistes pourraient résulter des taxa rares qui deviennent abondants avec le changement des conditions environnementales. Les choix méthodologiques associés à ces deux études dépendaient des programmes dans lesquels elles s'inséraient et étaient donc différents. Pour la première étude, qui faisait partie du programme EC2CO DIVAQUA (Responsable : Dr I. Mary), le choix a été fait, en raison du grand nombre d'échantillons (40 points d'échantillonnage), de confier la partie amplification par PCR et pyroséquençage a un prestataire externe. Pour la seconde étude, seule la partie pyroséquençage a été sous-traitée. Par conséquent, la marge de manœuvre était beaucoup plus importante dans ce dernier travail et nous avons pu associer étroitement expérimentation biologique et traitement bio-informatique.

3.2 Erreurs de séquençage : qualité et nettoyage

Si le développement des nouvelles techniques de séquençage a permis l'accès à la biosphère rare et ses patrons de diversité (Sogin et al., 2006, Galand et al., 2009), l'existence

de cette dernière demeure controversée du fait des biais associés au séquençage massif (Reeder et Knight 2009). En effet, les OTUs représentées par une ou peu de séquences sont souvent suspectées d'être artéfactuelles et sont de ce fait écartées de l'analyse. Afin de minimiser les erreurs de séquences et éviter une éventuelle inflation d'OTUs, nous avons appliqué un filtre qualité pour écarter les séquences : i) de longueur inférieure à 200 pb, ii) contenant une ou plusieurs bases indéterminées (N), iii) ne s'hybridant pas à 100% avec l'amorce *forward*, et iv) les séquences chimériques détectées par UCHIME (Edgar et al., 2011).

Les scores de qualité des bases nucléotidiques peuvent être utilisés de différentes manières. Par exemple, Nolte et al. (2010) calculent la qualité moyenne sur la longueur de la séquence et éliminent les séquences avec une qualité moyenne inférieure au seuil déterminé. Dans notre étude, les scores ont été utilisés tel que décrit dans PANGEA (Giongo et al., 2010), en coupant toutes les régions avec un score de qualité faible ; le score de qualité sur une région étant calculé en additionnant les différences entre le score de chaque position et le seuil fixé.

Dans l'article 3, le score de qualité choisi était de 27 (Kunin and Hugenholtz, 2010). Les différentes étapes de ce filtre nous ont alors conduit à éliminer environ 15% des séquences brutes résultant de l'amplification des *Archaea*. D'après (Huse et al., 2007), ces critères de qualité permettent de réduire les erreurs par base à $<0.2\%$ (environ une erreur tous les 450 nucléotides) sans nettoyage supplémentaire par les méthodes de « denoising » qui, au-delà du fait qu'elles nécessitent des ressources informatiques puissantes, ne sont pas encore universellement admises (Behnke et al., 2011). De plus, Vila-Costa et al. (2012) ont comparé une méthode de nettoyage proche de celle appliquée à notre jeu de données et une approche de denoising, et ont conclu à des résultats similaires.

La seconde étude se différencie de la première par l'ajout d'un témoin interne, correspondant à un produit de PCR d'un clone d'une séquence d'ADNr 18S de *Blastocystis hominis sous-type 4* (numéro d'accession : FJ666885), dans chaque échantillon à une proportion de 1%. Les séquences retrouvées dans chaque échantillon après amplification et séquençage nous ont permis de procéder à une normalisation quantitative des différents

échantillons, et dans un second temps de définir un score minimal de qualité ainsi qu'un seuil de similitude pour construire les OTUs. Sur les 348422 séquences brutes en sortie du séquenceur, 2946 étaient affiliées au témoin interne *B. hominis*. Parmi ces séquences, 24 ne possédaient pas de barcodes et ont été éliminées. Pour une date, l'ensemble des séquences ne formait qu'une seule OTU, qui plus est, ne correspondait pas au témoin interne *B. hominis* ; nous avons donc conclu à un problème sur cet échantillon et l'avons écarté. Sur les 2922 séquences de *B. hominis*, 1879 variants de la séquence initiale ont été identifiés. Près de 91.7% de ces variants n'étaient représentés que par une seule séquence. Enfin, 8.15% des séquences correspondant à *B. hominis* étaient identiques au standard interne initialement utilisé. L'origine de la variabilité observée entre les séquences de *B. hominis* ne peut être due qu'aux erreurs générées par l'amplification et/ou le séquençage. Afin de définir ces erreurs, une première analyse a été faite sur le type et la fréquence des mutations observées (Figure 3.1).

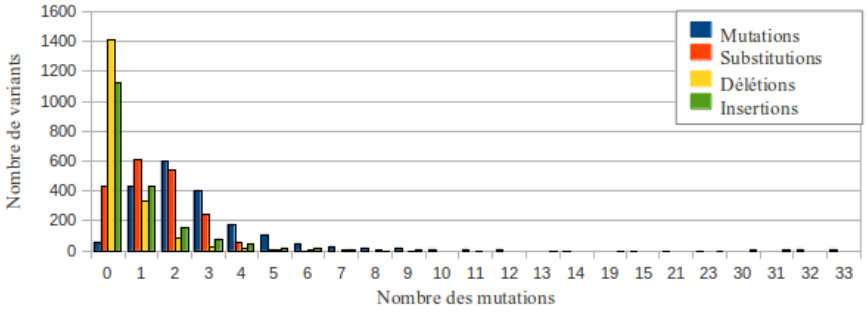


Figure 3.1. Types et fréquences des mutations par variant par rapport à la séquence de référence.
La fréquence des mutations correspond aux séquences comprenant au maximum une erreur, qui peut être une insertion, une délétion ou une substitution.

D'après la figure 3.1, il apparaît que les substitutions sont les principales erreurs observées, constituant environ 53% des mutations. L'analyse des scores de qualité associés aux bases nucléotidiques en fonction des types d'erreurs (figure 3.2.a) montre que seules les bases non résolues (Ns) possèdent de faibles scores (≈9) et peuvent être facilement détectées, tandis que les scores des insertions, bien qu'ils soient inférieurs à ceux des bases correctes restent élevés (≈27). Enfin, les substitutions possèdent des scores comparables à ceux des bases correctes, et ne peuvent donc être discriminées selon ce critère. La comparaison du taux d'erreur moyen (le pourcentage de positions avec une mutation) en fonction des scores de qualité (figure 3.2.b) montre que l'augmentation des scores de 20 à 29 n'a pas

d'effet sur la qualité des séquences puisque le taux d'erreur moyen ne change pas (≈0.4%). D'après cette première analyse des erreurs, il apparaît que les substitutions, indétectables par les scores de qualité, sont principalement à l'origine de la variabilité observée au sein des séquences de *B. hominis*. Ainsi, en plus d'un score de qualité minimale de 23 tel que utilisé par Nolte et al. (2010), les critères complémentaires classiques ont été également appliqués, conduisant à l'élimination du jeu de données de toutes les séquences avec au moins une base N, celles de taille inférieure à 200 pb, celles dont l'amorce F comporte des erreurs et les séquences chimériques. L'application de ces différents filtres a résulté en 1948 séquences nettoyées de *B. hominis* (dont 1263 variants). Sur le jeu de données global, environ 16% des séquences a été écarté.

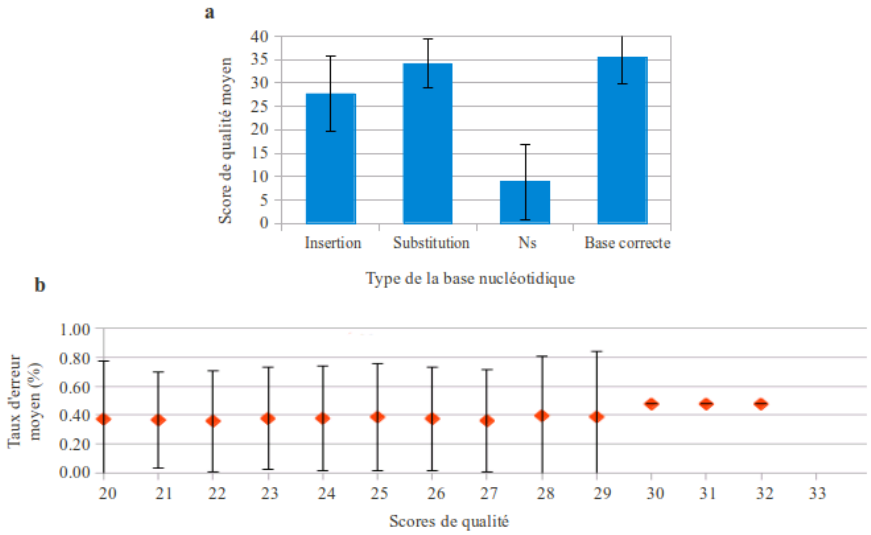


Figure 3.2. Les valeurs des scores de qualité en fonction du type de mutation (a) et du taux d'erreur (b).

3.3 Normalisation

Dans l'article 3, comme l'hétérogénéité du nombre de séquences nettoyées entre les différentes dates d'échantillonnage dans notre jeu de données (12 à 7910 séquences) risquait de biaiser les comparaisons de richesse et de diversité (Gihring et al., 2012), nous avons procédé à un ré-échantillonnage aléatoire de séquences de chaque point. Alors que le seuil de normalisation pose le problème de choix entre garder la plupart des échantillons avec de petits jeux de séquences ou garder peu d'échantillons avec de grands jeux de séquences, nous avons choisi dans cette étude de définir deux seuils : un seuil de 208 séquences nous

permettant de garder 68 échantillons pour un suivi régulier de la dynamique temporelle des communautés archéennes ; et un seuil de 488 séquences, écartant ainsi 25 échantillons mais offrant une profondeur de séquençage plus importante pour détecter la fraction rare.

Dans la deuxième étude, les tailles des jeux de séquences des différents échantillons variaient de 7353 à 18967 séquences, alors que le nombre des séquences de *B. hominis* allait de 23 à 654. Compte tenue de la proportion du standard interne avant (1% dans chaque échantillon), et après séquençage, nous avons pu calculer pour chaque échantillon, un facteur de correction des effectifs.

3.4 Seuils de similitude et OTUs rares

Pour la définition du seuil de clusterisation des *Archaea*, nous nous sommes basés sur les travaux de Kim et al. (2011) et choisi une similitude de 97% pour la région hypervariable v3-v5, zone ciblée lors de cette étude. En revanche pour la deuxième étude, nous pouvions nous appuyer sur le standard interne pour déterminer un seuil de clustérisation correspondant aux conditions expérimentales, à la méthode de clustérisation utilisée et aux organismes étudiés. Plusieurs seuils ont été appliqués aux 1948 séquences nettoyées de *B. hominis*. L'hypothèse nulle de ce test est que le meilleur seuil est celui qui regroupe toutes les séquences dans la même OTU, étant donné qu'elles proviennent toutes du même clone. Trois seuils de similitude ont été testés : celui de 97% traditionnellement utilisé pour la définition des OTUs, celui de 95% défini pour les protistes par Caron et al. (2009), et un seuil intermédiaire de 96% (Tableau 3.1).

D'après le tableau 3.1, il apparaît que le nombre des OTUs générées diminue lorsque le seuil de clusterisation passe de 97% et 95% quelle que soit l'approche utilisée. Cependant, au seuil de 95% UCLUST donne la meilleure approximation puisqu'il génère deux OTUs dont un singleton. Le pourcentage d'identité minimal séparant ce singleton résiduel de l'OTU majoritaire était de 26%, l'obtention d'une seule OTU nécessitant donc une clusterisation à 74%. Au seuil de 95%, Mothur génère 53, 77 et 122 OTUs avec les algorithmes NN «Nearest Neighbor», AN «Average Neighbor», et FN «Furthest Neighbor» respectivement dont de nombreux singletons (NN : 28, AN : 41 et FN : 55). Ainsi, notre procédure de nettoyage associée à une clusterisation avec UCLUST au seuil de 95% nous permet

Table 3.1. Nombre et structure des OTUs générées par différentes méthodes à différents seuils de similitude pour les 1948 séquences nettoyées de *B. hominis*.

Méthode	UCLUST			Mothur FN			Mothur AN			Mothur NN		
Similitude	95	96	97	95	96	97	95	96	97	95	96	97
Nbr OTUs	2	4	9	122	154	208	77	104	149	53	66	98
Nbr Singletons	1	1	1	55	73	113	41	58	98	28	37	66
Nbr seq max par OTU	1947	1826	1788	1194	1131	1077	1285	1250	1233	1308	1305	1294

UCLUST implémente une méthode heuristique, alors que Mothur calcule une matrice de distance par alignement multiple, avant de générer les OTUs en comparant avec le seuil de similitude soit la distance moyenne entre les séquences d'un OTU (AN pour average neighbor) ; soit la distance avec le voisin le plus éloigné (FN pour furthest neighbor) ou le voisin le plus proche (NN pour nearest neighbor).

vraisemblablement d'avoir la meilleure estimation de la richesse réelle des séquences de *B. hominis*. Enfin, alors que la clusterisation à 95% avec UCLUST des séquences brutes de *B. hominis* génère 23 OTUs dont un contenant 1847 séquences, notre procédure de nettoyage nous a permis de réduire l'inflation des OTUs et d'avoir une meilleure estimation de la richesse réelle des séquences de *B. hominis*. Afin de comparer le regroupement en OTU des séquences affiliées à *B. hominis* par rapport à leur positionnement phylogénétique, une phylogénie contenant les 1948 séquences nettoyées en plus de séquences de référence a été générée (Figure 3.3). D'après cette figure, il apparaît que les séquences forment un groupe monophylétique y compris la séquence du singleton (22 G2SRB0Y01AU4G3).

La définition de la fraction rare reste encore relativement subjective dans la mesure où les seuils utilisés sont différents selon les études (e.g., <1% d'après Campbell et al. (2011) ; <0.1% d'après Pedrós-Alió (2006) ; Sogin et al. (2006) et Fuhrman (2009) ou encore <0.05% d'après Vila-Costa et al. (2012)), et qu'ils dépendent de la profondeur de séquençage et de la diversité initiale du milieu échantillonné. Lors de notre travail, le seuil délimitant les rares des abondants pour les *Archaea* a été choisi sur la base du seuil de normalisation (488), et ont été considérées comme OTUs rares toutes celles contenant au maximum 0.2% des séquences. Pour la seconde étude, nous avons déterminé trois classes, celle des rares correspondants à ($\leq$0.01%), celle des abondants à ($\geq$1%) et les autres OTUs étant classées comme intermédiaires.

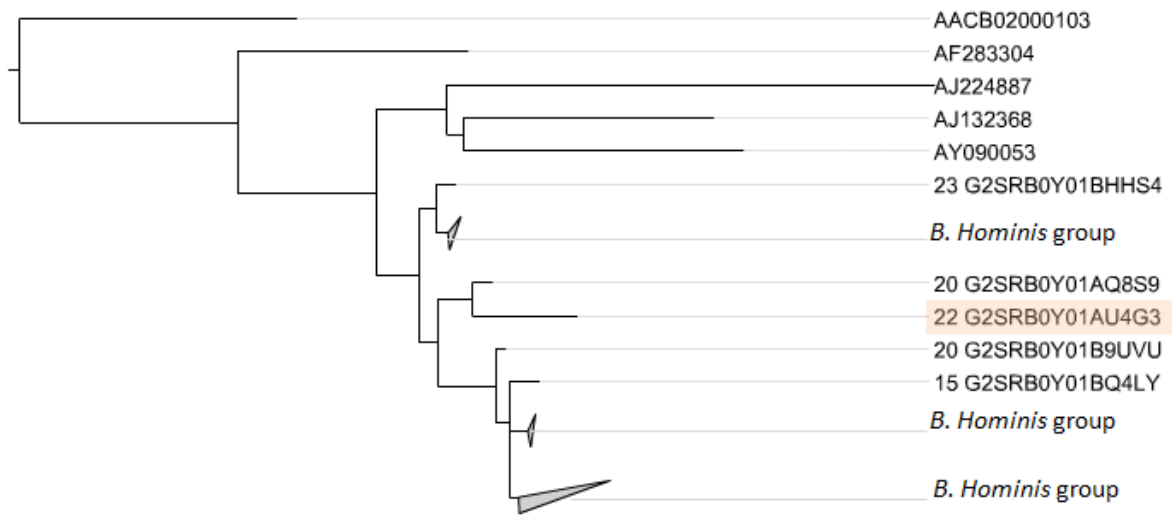


Figure 3.3. La phylogénie des 1948 séquences de *B. hominis* avec des séquences de référence.
Les 1948 séquences de *B. hominis* constituent un groupe monophylétique incluant la séquence (22 G2SRB0Y01AU4G3).

3.5 Etude des clades

L'utilisation de séquences de taille réduite issues du pyroséquençage pour construire des arbres phylogénétiques pourrait être problématique pour la robustesse des phylogénies inférées. Cependant, des études récentes ont montré que les méthodes phylogénétiques étaient fiables même avec ce type de données (Jeraldo et al., 2011, Ragan-Kelley et al., 2012). Ceci est conforté par nos résultats qui montrent que, l'affiliation phylogénétique de nos OTUs restitue des clades d'*Archaea* et de protistes identiques à ceux définis dans la littérature par du clonage/séquençage (Lefranc et al., 2005, Galand et al., 2010), ce qui indique que des séquences courtes (≈450 pb) peuvent être aussi informatives que des séquences complètes, du moins pour la délimitation des groupes monophylétiques. Autres que les clades déjà définis et comportant à la fois des OTUs abondantes et des OTUs rares, les phylogénies générées nous ont permis également de décrire deux nouveaux clades, constitués cette fois exclusivement d'OTUs d'*Archaea* rares. Ces derniers, en plus de former des groupes monophylétiques et distincts, possèdent des dynamiques d'activité différentes qui se détachent de celles des clades abondants, et semblent répondre aux changements des conditions environnementales, leur activité étant plus importante en hiver.

Notre approche phylogénétique couplée à l'utilisation des nouvelles techniques de séquençage nous a permis de mettre en évidence de nouveaux clades d'*Archaea* constitués uniquement d'OTUs rares. L'étude de cette diversité phylogénétique à la lumière des données environnementales montre que ces OTUs suivent une dynamique saisonnière et qu'ils reflètent une réalité écologique, minimisant ainsi l'hypothèse d'artefacts méthodologiques dans le cadre de cette étude. Dans la seconde publication, l'utilisation d'un standard interne nous a permis de normaliser les effectifs des librairies des échantillons sans effectuer un ré-échantillonnage. En effet, la normalisation "classique" ramène tous les échantillons aux mêmes effectifs, ce qui peut dissimuler des différences d'abondance. D'un autre côté, l'étude des séquences du standard interne nous a permis de définir un filtre de qualité et un seuil de clusterisation qui ne surestiment pas la richesse et les espèces rares. Nous avons ainsi pu mettre en évidence d'importantes variations quantitatives à court terme à l'échelle de la communauté picoeucaryotique.

Article 3

Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters

Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters

Mylène Hugoni[a,b,1], Najwa Taib[a,b,1], Didier Debroas[a,b], Isabelle Domaizon[c], Isabelle Jouan Dufournel[a,b], Gisèle Bronner[a,b], Ian Salter[d,e], Hélène Agogué[f], Isabelle Mary[a,b,2], and Pierre E. Galand[d,g]

[a]Laboratoire "Microorganismes: Génome et Environnement," Clermont University, Université Blaise Pascal, F-63000 Clermont-Ferrand, France; [b]Laboratoire Microorganismes, Génome et Environnement, Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche (UMR) 6023, F-63171 Aubière, France; [c]Institut National de la Recherche Agronomique, UMR 42 Centre Alpin de Recherche sur les Réseaux Trophiques et Ecosystèmes Limniques, F-74200 Thonon les Bains, France; [d]Université Pierre et Marie Curie–Paris 6, UMR 8222, Laboratoire d'Ecogéochimie des Environments Benthiques (LECOB), UMR 7621, Laboratoire d'Océanographie Microbienne (LOMIC), Observatoire Océanologique, F-66650 Banyuls-sur-Mer, France; [e]CNRS, UMR 7621, LOMIC, Observatoire Océanologique, F-66650 Banyuls-sur-Mer, France; [f]Littoral, Environnement et Sociétés, UMR 7266, CNRS, University of La Rochelle, 17000 La Rochelle, France; and [g]CNRS, UMR 8222, LECOB, Observatoire Océanologique, F-66650 Banyuls-sur-Mer, France

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved March 1, 2013 (received for review September 28, 2012)

Marine Archaea are important players among microbial plankton and significantly contribute to biogeochemical cycles, but details regarding their community structure and long-term seasonal activity and dynamics remain largely unexplored. In this study, we monitored the interannual archaeal community composition of abundant and rare biospheres in northwestern Mediterranean Sea surface waters by pyrosequencing 16S rDNA and rRNA. A detailed analysis of the rare biosphere structure showed that the rare archaeal community was composed of three distinct fractions. One contained the rare Archaea that became abundant at different times within the same ecosystem; these cells were typically not dormant, and we hypothesize that they represent a local seed bank that is specific and essential for ecosystem functioning through cycling seasonal environmental conditions. The second fraction contained cells that were uncommon in public databases and not active, consisting of aliens to the studied ecosystem and representing a nonlocal seed bank of potential colonizers. The third fraction contained Archaea that were always rare but actively growing; their affiliation and seasonal dynamics were similar to the abundant microbes and could not be considered a seed bank. We also showed that the major archaeal groups, Thaumarchaeota marine group I and Euryarchaeota group II.B in winter and Euryarchaeota group II.A in summer, contained different ecotypes with varying activities. Our findings suggest that archaeal diversity could be associated with distinct metabolisms or life strategies, and that the rare archaeal biosphere is composed of a complex assortment of organisms with distinct histories that affect their potential for growth.

ong-term dynamic | dormancy | taxonomic diversity | microbial observatory | Somlit

The seasonal dynamics of marine microorganisms have traditionally been studied at the DNA level (1, 2), but recent studies have shown the importance of differentiating the active communities from the total communities (3–5). One method to explore an aspect of activity (i.e., the growth rate for specific taxa) is to investigate microbial communities with both 16S rRNA and 16S rDNA (6-8). The use of the 16S rRNA-to-rDNA sequence ratio as an index of microbial growth has revealed a generally positive correlation between abundance and activity in coastal surface bacterial communities (4, 9). However, abundant microbes are not always the most active (3), even though they contribute greatly to ecosystem functioning. An important finding is that growth can be detected among low-abundance taxa, also known as the rare biosphere (4, 7), which was first defined with the development of new sequencing technologies, allowing a deep coverage of the diversity of natural communities (10). Rare taxa have been hypothesized to consist of dormant microorganisms (or a seed bank) that could potentially be resuscitated under different environmental conditions (11). However, the discoveries that the rare biosphere had a biogeography (12), and that a significant portion of the rare community was active (4, 7), with growth rates that decreased as abundance increased (4), suggest that the rare biosphere is not solely a dormant seed bank (13). A rare biosphere has been detected within the domain Archaea (12), and although we have begun to gain insights into the dominant archaeal phylotypes, the community structure of the rare Archaea remains largely uncharacterized.

Marine planktonic Archaea have been recently recognized as main drivers of the aerobic ammonia oxidation in many aquatic ecosystems, suggesting an important role in the nitrogen cycle (14–16). They have traditionally been described as spanning three major groups: Thaumarchaeota marine group (MG) I, which is more abundant in meso- and bathypelagic waters (17–19), Euryarchaeota MGII, which is more abundant in surface waters, and Euryarchaeota MGIII, which is restricted to deeper waters (20, 21). The diversity of Archaea is, however, much more complex; for instance, MGI appears to have distinct clusters segregated according to depth and location (22). A recent metagenomic characterization of MGI from north Atlantic coastal surface waters also suggested the presence of at least two dominant environmental populations that are divergent from each other (23). The presence of at least two clusters was also demonstrated in the Mediterranean Sea (24) and corresponded to groups previously detected in different oceanic provinces (20). Whether this taxonomic diversity corresponds to distinct ecotypes, i.e., groups of microorganisms playing distinct ecological roles and belonging to genetically cohesive and irreversibly separate evolutionary lineages (25), is not known because the relationship among archaeal activity, environmental conditions, and sequence abundance has never been studied. Moreover, the ecological control of archaeal diversity patterns over long time scales remains poorly understood (24).

By monitoring surface archaeal communities in monthly intervals during a 3.5-y period at the Banyuls-sur-Mer Bay Microbial Observatory, a site representative of the coastal northwest Mediterranean Sea, we aimed to describe the structure of the rare archaeal

Author contributions: I.D., H.A., I.M., and P.E.G. designed research; M.H. and N.T. performed research; M.H., N.T., D.D., I.S., and I.M. contributed new reagents/analytic tools; M.H., N.T., I.J.D., G.B., and P.E.G. analyzed data; M.H., N.T., I.M., and P.E.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The pyrosequencing data reported in this paper has been deposited in the Dryad database, http://datadryad.org (doi no. 10.5061/dryad.q5903).

[1]M.H. and N.T. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: isabelle.mary@univ-bpclermont.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1216863110/-/DCSupplemental.

biosphere by testing whether it is composed of a seed bank of dormant cells or represents microorganisms with high growth rates. By targeting both 16S rRNA and rDNA, we also verified whether different archaeal clusters represent distinct ecotypes and assessed the seasonal activity dynamics of marine Archaea.

Results

Rare and Abundant Phylotypes. We used pyrosequencing to follow changes in the community structure, relative sequence abundance, and potential activity of Archaea over time. A total of 351 operational taxonomic units (OTUs) were retrieved in the 16S rDNA dataset, representing a total of 65,833 sequences. Seventeen OTUs were abundant (>1% and occurred in more than one sample) and contained 97% of all of the sequences. The 16S rRNA dataset was composed of 52,181 sequences consisting of 348 OTUs. Rarefaction curves for both 16S rDNA and 16S rRNA indicated that, in most cases, the sequencing depth captured the diversity present in the natural archaeal community (Fig. S1).

Only two OTUs were always abundant (OTU 2 affiliated with MGI, and OTU 13 affiliated with MGII.B), whereas the remaining 15 abundant OTUs were rare in some samples (≤0.2% of the sequences in a sample). Typically, OTUs abundant in winter became rare in summer and vice versa. All the abundant OTUs were active when they were abundant (Fig. 1*A*), and the plot of the 16S rDNA against 16S rRNA OTU frequencies had an intercept at zero and showed a high correlation between 16S rRNA and 16S rDNA (Kendall nonparametric $\tau = 0.7$; $P < 0.001$; $n = 224$). However, 16S rRNA and 16S rDNA were more poorly correlated when the abundant OTUs became rare ($\tau = 0.3$; $P < 0.001$; $n = 72$), with some OTUs showing high activity (16S rRNA/rDNA ratio >1) whereas others had low or no activity (16S rRNA/rDNA ratio <1; Fig. 1*B*).

All OTUs were compared with the entire SILVA database to ascertain if they were globally common (i.e., a high similarity to reference sequences) or uncommon (i.e., a low similarity). The abundant DNA OTUs were common, with an average 98% sequence similarity to the public database sequences (Fig. 2*A*). The always rare DNA also contained a group of common OTUs (96% similarity), but, notably, half the OTUs were uncommon, with only 84% identity to the public reference sequences (Fig. 2*B*). The abundant and always-rare RNA OTUs were common (98% and 96% sequence identity, respectively; Fig. 2) for the active fraction of the community, and the absence of uncommon OTUs in the rare 16S rRNA fraction indicates that the uncommon OTUs were never active. The low similarity to the SILVA database displayed by the uncommon rare OTUs (84% identity) suggests that they originated from undersampled ecosystems, not well covered by the public database. The closest relatives to the uncommon rare OTUs belonged to the Euryarchaeota Deep Hydrothermal Vent Euryarchaeotic Group 6 (DHVEG-6), pMC1, and South African Gold Mine Euryarchaeotic Group-1 (SAGMEG-1) clusters, which are frequently detected in deep marine sediments (26). In contrast, the rare but common OTUs were identified as MGI and MGII.

Archaeal Community Structure, Dynamics, and Activity. The OTUs abundance followed a log-series distribution for the 16S rDNA and rRNA datasets (Fig. S2 *A* and *B*), with most OTUs included in the first octaves (i.e., species characterized by a low number of reads). The abundant OTUs belonged mostly to MGI and MGII.A and MGII.B (Fig. S2*C*) but also to MGIII. In the active fraction (the 16S rRNA dataset), the major taxonomic groups (MGI, MGII.A, and MGII.B) represented ∼93% of the reads (Fig. S2*C*). Interestingly, some abundant OTUs in the 16S rDNA dataset were less represented in the 16S rRNA dataset, for example, OTUs 2 and 9 affiliated with MGI, suggesting a weak activity. In contrast, some abundant OTUs were also very active, as shown by a greater relative abundance of 16S rRNA, such as OTU 28 affiliated with MGI (Fig. S2*C*).

The MGI sequences followed a seasonal pattern and were more abundant during winter (Fig. 3*A*). The MGI 16S rRNA dynamics showed the same trend as that for 16S rDNA, suggesting metabolically active communities. Our analysis showed that the MGI OTUs fell into four different clusters: A, B, C, and D (Fig. S3*A*). Most of the OTUs were affiliated with MGI.B, followed by MGI.A, which is closely related to *Nitrosopumilus maritimus*. These two clusters comprised all the abundant MGI OTUs. Interestingly, MGI.A and MGI.B sequences exhibited alternative patterns of 16S rDNA and rRNA representation. The MGI.A sequences outnumbered the MGI.B sequences in the 16S rDNA dataset (approximately two times more), whereas the opposite was observed for the 16S rRNA dataset (Fig. 4). This result suggested that MGI.B was much more active than MGI.A, even though it was not the most abundant in the ecosystem. The rare OTUs belonged to MGI.C, which is affiliated with sequences retrieved from deep waters (20) and distantly related to *Cenarchaeum symbiosum* and to MGI.D. This cluster was distinct from the others (89–92% similarity) and emerged earlier in the phylogeny. MGI.C was active when present, whereas MGI.D was not always active when present.

The MGII.A sequence abundance showed marked differences between seasons, with the highest relative abundance during the summer period (from May to October) and the lowest during the winter months (Fig. 3*B*). The 16S rRNA dataset revealed a similar seasonal pattern of activity. In contrast, MGII.B dominated in abundance and activity during winter, with the highest relative abundance in February and recurrent peaks each year (Fig. 3*C*). MGII.A was more active than MGII.B, consistent with its higher relative abundance (Fig. 3 *B* and *C*). Euryarchaeota MGII.A was

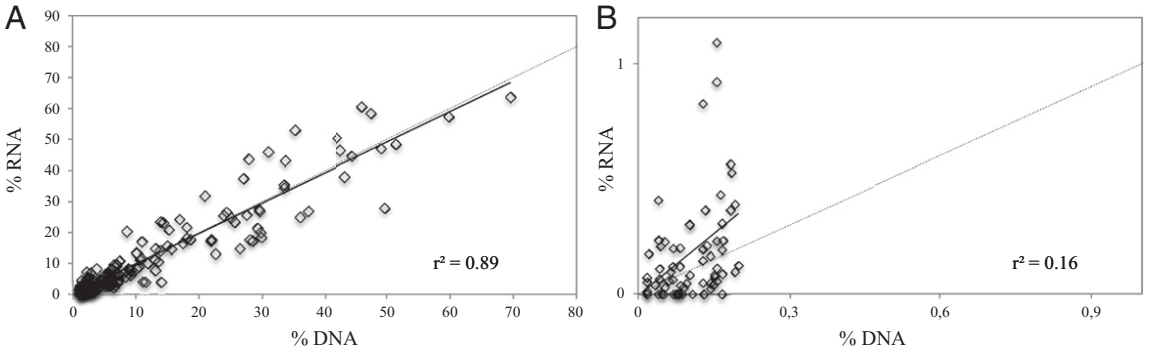


Fig. 1. 16S rDNA against 16S rRNA OTUs frequencies for abundant OTUs when they are abundant (*A*) and when they become rare (*B*). The RNA and DNA frequencies are plotted against each other for all abundant OTUs and all time points. The black line represents the regression, and the dotted line is the 1:1 line.

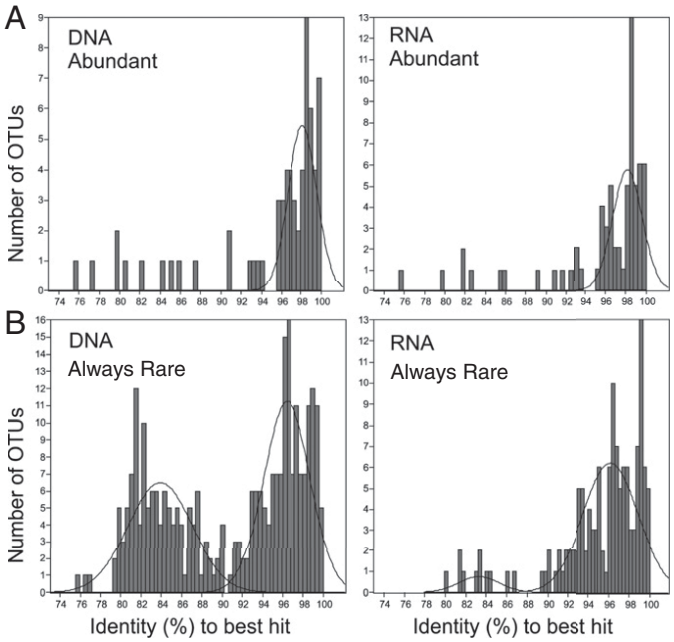


Fig. 2. Distribution of the percent identity from a comparison between public database sequences (SILVA) and abundant 16S rDNA and rRNA sequences (*A*) and always-rare 16S rDNA and rRNA sequences (*B*). The data are fitted to a model of normal distributions (black lines) that identifies groups of OTUs as common (i.e., a high percentage identity) or uncommon (i.e., a low percentage identity).

separated into two previously described (20) main subclusters (M and K; Fig. S4), and most of the sequences belonged to subcluster M, which was also the most active (Fig. S5*A*). The subcluster M activity pattern was different from that of subcluster K (Fig. S5). Most MGII.B sequences and activity were affiliated with the WHARN subcluster (Figs. S5*B* and S6) that corresponds to phylotypes II-CC, which are widely distributed in surface waters of various oceanic provinces (20). Other Euryarchaeota were affiliated with the MGIII and the RC-V cluster and with methanogenic lineages (Fig. S6).

Less abundant groups, including OTUs affiliated with MGIII, were also present and active during winter but were also detected in July 2008 and 2009, together with reduced activity. The Miscellaneous Euryarchaeotic Group (MEG) and DHVEG-6 did not present seasonal patterns of relative abundance and activity.

The canonical correspondence analysis plot (*SI Materials and Methods*) showed a clear difference between the activity of the two MGII clusters (Fig. S7): MGII.A appeared as a summer community associated mainly with temperature, whereas the activity of MGII.B was related to such winter features as nitrite, nitrate, and oxygen. These winter features also characterized the activity of MGI overall, and there were fewer differences between the different MGI clusters when considering the parameters followed in the present study. Contrary to MGII, MGI clusters were discriminated according the second axis, which was positively correlated with phosphate (Fig. S7).

Discussion

Our long-term study of archaeal dynamics and activity in surface Mediterranean waters showed that rare Archaea were heterogeneous in their pattern of seasonal activity and phylogenetic affiliation. We propose that the rare archaeal biosphere could be divided into three different fractions classified as follows: the local seed bank, the nonlocal seed bank (or the alien colonizers), and the active-but-always-rare fraction.

The local seed bank represented Archaea that were rare but became abundant at certain times. When abundant, their 16S rDNA and 16S rRNA sequences were closely correlated, indicating that these OTUs were also active. Scatter plots of 16S rRNA vs. rDNA yielded an intercept at zero, suggesting that growth rates were constant as abundance varied (4). However, when these OTUs became rare, their 16S rDNA and rRNA sequences were poorly correlated, which, according to a described model (4), indicates increasing or decreasing growth rates as abundance decreases. Such variable activity suggests changing growth rates, possibly reflecting differences in the metabolic state of the cells as they cycle between abundant and rare fractions. Contrasting activity levels among rare microbes have been reported recently for Bacteria in a coastal system (4) and in lakes (7). Within the context of our seasonal study, the observations could indicate that these rare microorganisms are able to react to seasonal fluctuations of environmental conditions. Moreover, these Archaea could not be considered as being typically dormant cells because some of them lacked dormancy stages (i.e., were always active) and others had only short ones. We propose that this local seed bank maintains sufficient metabolic diversity to react to fluctuating environmental conditions.

The second fraction contained rare Archaea that were uncommon and always inactive in the northwestern Mediterranean Sea. They were aliens to the studied pelagic ecosystem, and their low similarity to database sequences indicates that they may originate from undersampled ecosystems, such as deep marine sediments. This nonlocal seed bank may be dispersed by such episodic events as river flooding, strong storms, or even atmospheric deposition. It is possible that these microorganisms may never grow in the water column as a result of a requirement for very different conditions to those found in the pelagic environment. This fraction of the rare archaeal biosphere could be on its way to extinction (13); alternatively, it may have the ability to

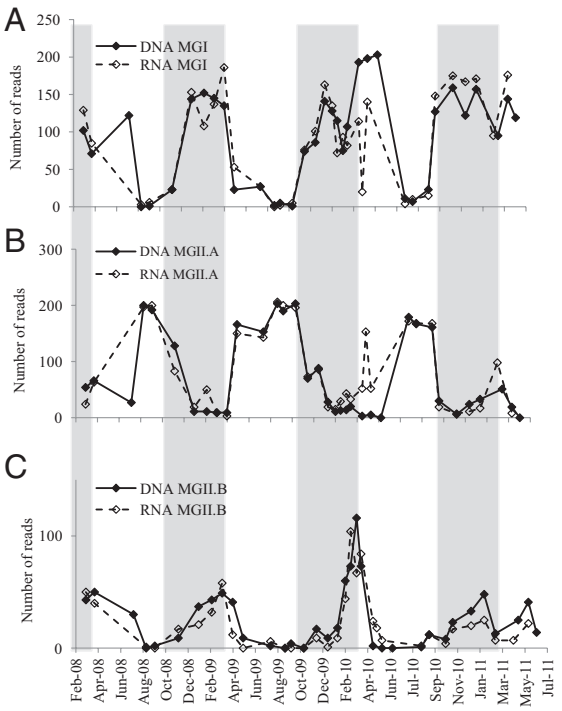


Fig. 3. Seasonal dynamics of the most abundant taxonomic groups in both the 16S rDNA and rRNA sequence datasets: (*A*) MGI, (*B*) MGII.A, and (*C*) MGII.B. Winter months are shown in gray; summer months are shown in white.

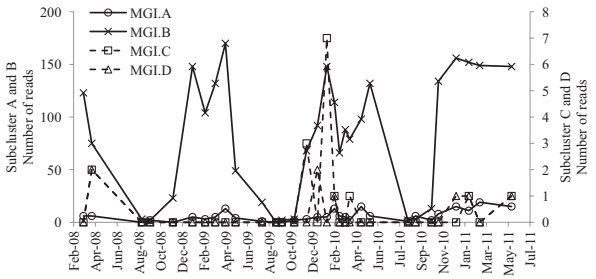


Fig. 4. Seasonal dynamics of active (16S rRNA dataset) MGI clusters: MGI.A, MGI.B, MGI.C, and MGI.D. Winter months are shown in gray; summer months are shown in white.

survive in a dormant stage representing a pool of potential colonizers of very different ecosystems.

Finally, the third fraction of rare Archaea was represented by cells that were always rare but actively growing, and their phylogenetic affiliation and seasonal dynamics were similar to those of the abundant Archaea. Although their activity suggests that these cells are not dormant, the fact that they never made the transition from rare to abundant indicates that they are not operating as a seed bank under the range of environmental conditions encountered during this study. Always rare but actively growing populations may be maintained at a low abundance because they are less efficient than others when competing for resources, or they may have a different life strategy that promotes rarity to minimize predation (11). Alternatively, these groups may be more susceptible to specific viral attack (27).

Although MGI was overall more abundant and active in winter, we found differences in activity between the clusters. Four MGI clusters were identified through our analysis: MGI.A (or I-α), MGI.B (or I-β), and MGI.C (I-γ) were first described in a large-scale study based on terminal restriction fragment length polymorphism (T-RFLP) (20), but only MGI.A and MGI.B were reported in a previous study conducted in the Mediterranean Sea (24). The similar delineation of MGI.A and MGI.B by cloning/sequencing (24) and pyrosequencing confirmed that shorter pyrosequencing reads may be as informative as near full-length sequences, and thus suitable for building phylogenies (28, 29). Interestingly, the MGI.A sequences were more abundant within the 16S rDNA fraction, whereas MGI.B was more abundant within the 16S rRNA fraction, suggesting that the most abundant MGI cluster (i.e., MGI.A) was not the most active. The rarest clusters, MGI.C and MGI.D, also presented different patterns of activity: MGI.C was active when present, yet MGI.D was not always active when present. Both of these rare clusters were active in winter, showing similar responses to such environmental conditions as oxygen, nitrite, nitrate, or Chlorophyll *a* (Chl*a*) content. However, there was a mismatch between their peaks of maximum activity. The variable activity levels for the different MGI clusters could suggest the presence of MGI clusters adapted to different niches. The clusters' recurrent seasonal activity patterns concomitant with reproducible environmental conditions may indicate an ecological specialization and predictable population–environment linkage, comforting the idea of separate ecotypes, as previously shown for Vibrionaceae populations (30). Although we could not precisely define niches in the present study, we hypothesize that the MGI clusters represent ecotypes that correspond to distinct metabolisms, as previously demonstrated for Bacteria (4). In fact, closely related SAR11 phylotypes were recently shown to have very different growth rates (4), which could correspond to differences in metabolism, such as those observed between the SAR11 ecotypes for phosphorus acquisition or glucose utilization (31, 32). Different metabolisms could thus allow the MGI ecotypes to occupy a variety of niches and explain a global ecological success that is similar to that of SAR11 (33).

The MGI seasonal dynamics, illustrated by a higher relative abundance during winter, correlated with the seasonality of such surface water environmental parameters as nitrite and nitrate concentrations. MGI is thought to play an important role in the marine nitrogen cycle by oxidizing ammonia to nitrite (34, 35). In our survey, most of the 16S rDNA sequences belonged to MGI.A, which is closely related to the cultivated planktonic marine Archaea *N. maritimus* (36), an autotrophic ammonia-oxidizer that produces nitrite, suggesting the potential for ammonia oxidation of the MGI found in these waters. This hypothesis is supported by the MGI winter peaks that coincided with an increase in the nitrite and nitrate concentrations, followed by a decrease in ammonia. However, because phytoplankton can also release nitrite (37), the exact origin of winter nitrite in this ecosystem cannot be conclusively determined. Moreover, MGI.C was distantly related to *C. symbiosum*, which is able to use urea as energy and carbon source (38). As archaeal genes for urea utilization have been detected in the marine environment (39–41), we can also speculate that urea could be used as an alternative source of nitrogen and carbon by the MGI clusters that are not closely related to *N. maritimus*, which does not possess genes for urea utilization (42). The decrease in MGI abundance and activity was coincident with an increase in Chl*a*, as previously shown in marine and freshwater ecosystems (43–45). This finding suggests two hypotheses: a limitation of MGI abundance by organic material excreted by phototrophic primary production (36) or a competition with phytoplankton for ammonium that may be unfavorable to MGI (46).

Euryarchaeota were present year-round, but the opposite seasonal dynamics of MGII.A and MGII.B suggests the presence of different ecotypes. The predominance of MGII.A in summer corresponds to a season of generally low abundance of MGII and Archaea in the northwestern Mediterranean Sea (24). The winter peak of MGII.B corresponds to the highest archaeal and MGI abundances (24), suggesting separate niches for the two MGII clusters in surface waters. Their abundance and activity dynamics might be affected by competition for resources with other organisms and could reflect the development of different strategies to improve their metabolic potential. For instance, genes encoding proteorhodopsin have been found in members of the Euryarchaeota MGII.A in the surface waters of the North Pacific (47), and recent studies showed a single copy of the proteorhodopsin (*pop*) gene in the reconstruction of a coastal MGII.A genome (48). In our study, the MGII.A sequences were only 90% similar to the proteorhodopsin clade from a previous work (47), but closer to the sequence reported in another (48) (96% similarity). We therefore hypothesize that phototrophic metabolism could be present in the Mediterranean Sea: MGII.A could use light as an energy source, explaining the summer peaks of abundance and activity, as irradiance is more important in this season. In contrast, the MGII.B distribution in the rRNA datasets correlated with nitrogen compounds, as with MGI.

The results based on the 16S rRNA/rDNA ratios could have been affected by the 16S rDNA copy number per genome. However, to our knowledge, all available complete genomes of mesophilic Archaea, including representatives from Euryarchaeota and Thaumarchaeota showed only one copy of 16S rDNA: thus, we assume that this is also the case in the natural communities. To assess how well our sequence data could represent the community abundance, we compared the pyrosequencing quantification to the metagenomic data obtained in September 2010 from the same sampling station through the Global Ocean Sampling project. The two methods showed a strong dominance of Euryarchaeota (95% and 88% for the pyrosequencing and metagenomic data, respectively) vs. Thaumarchaeota, suggesting no major primer bias at the phylum level in our approach. A

significant correlation has also been found between the abundance estimated by quantitative methods and pyrosequencing for Bacteria (4, 6, 9) and by sequencing approaches for MGI in the northwestern Mediterranean Sea (24). We therefore hypothesize that the relative sequence abundance measured in this study was comparable to the cell abundance dynamics.

In summary, this study clearly showed that the rare biosphere could not solely be characterized as a seed bank of dormant cells; rather, it is a complex association of indigenous and itinerant cell types with contrasted origins and fate that contribute to microbial interaction networks and metabolic processes in the environment. Our phylogenetic affiliation suggested that the diversity found within the environmental clusters of Archaea may correspond to different activity levels or growth rates, thus possibly illustrating different metabolism and life strategies. Our results show that we need to rethink our view of how abundant and rare microbes contribute to ecosystem processes.

Materials and Methods

Sampling and Environmental Parameters. Surface seawater (3 m) was collected monthly from March 2008 to June 2011 (40 samples) by using a 10-L Niskin bottle at the Service d'Observation du Laboratoire Arago station (42°31′N, 03°11′E) in the Bay of Banyuls-sur-Mer in France. The water was kept in 10-L high density polyethylene carboys in the dark until being processed in the laboratory (within 1.5 h). A subsample of 5 L was prefiltered through 3-μm pore-size polycarbonate filters (Millipore), and the microbial biomass was collected on 0.22-μm pore-size GV Sterivex cartridges (Millipore) and stored at −80 °C until nucleic acid extraction. The physicochemical parameters (Fig. S8) were provided by the Service d'Observation en Milieu Littoral (www.domino.u-bordeaux.fr/somlit_national).

The water sample used for the metagenomic analysis was collected at 3 m depth on 28 September 2010 as part of the J. Craig Venter Institute European Sampling Expedition following a protocol previously published (49). Annotation of the metagenomic data were performed through the J. Craig Venter Institute metagenomics analysis pipeline (San Diego) (50).

Nucleic Acid Extraction and Pyrosequencing. The nucleic acid extraction method was modified from Lami et al. (8) by using a combination of mechanic and enzymatic cell lysis applied directly to Sterivex cartridges, followed by extraction by using the AllPrep DNA/RNA kit (Qiagen). The RNA samples were tested for the presence of contaminating genomic DNA by PCR and then reverse-transcribed with random primers using the SuperScript III Reverse Transcriptase kit (Invitrogen). The amplification of the V3–V5 region of the 16S rRNA gene was performed by Research and Testing Laboratory (Lubbock, TX) with universal archaeal primers Arch349F (CCC TAC GGG GTG CAS CAG) and Arch806R (GGA CTA CVS GGG TAT CTA AT) (51), followed by pyrosequencing by using a Roche 454 GS-FLX system with titanium chemistry.

Bioinformatic Analysis and Statistics. The pyrosequencing data produced from the 80 samples (16S rDNA and 16S rRNA) represented 477,589 raw sequences. All sequences were checked against the following quality criteria: (*i*) no Ns; (*ii*) quality score ≥27 according to PANGEA trimming (52); (*iii*) a minimum sequence length of 200 bp; (*iv*) no sequencing error in the forward primer; and (*v*) no chimeras [checked with UCHIME (53)]. The quality filtering step eliminated ∼15% of all sequences (1.6% were chimeras). The remaining reads were clustered using USEARCH (54) at a 97% similarity threshold (55). For the taxonomic affiliation, we constructed a dedicated archaeal database based on the SSURef 108 database of the SILVA project (56) and added annotated reference sequences from the Mediterranean Sea (24). The process was automated by PANAM (http://code.google.com/p/panam-phylogenetic-annotation/downloads/list) that constructs phylogenetic trees for taxonomic annotation (57) as detailed in *SI Materials and Methods*. After that step, all sequences affiliated to Bacteria were removed from the data set, leaving a total of 65,833 archaeal sequences for the 16S rDNA dataset and 52,181 sequences for the 16S rRNA dataset (Table S1). Phylogenetic trees containing only the main taxonomic groups detected by PANAM (MGI Thaumarchaeota, MGII.A and MGII.B Euryarchaeota), and environmental OTUs affiliated with those groups, are included as Figs. S3, S4, and S6.

For the analysis of the seasonal dynamics, the 16S rDNA and 16S rRNA samples were randomly resampled down to 208 sequences by using DaisyChopper (www.genomics.ceh.ac.uk/GeneSwytch/). We chose to resample down to a relatively low number of sequences to retain the largest possible number of samples; a total of 12 samples were discarded because of a low number of sequences (< 208). However, for the analysis of the rare biosphere, a deeper sequencing effort was needed to define the rare Archaea, and only samples with >488 sequences were retained (55 samples). To verify if the different sampling cutoff could bias our analysis, we compared the seasonal dynamics based on 208 sequences per samples to that based on 488 sequences. The two results were similar for the major groups, as, for example, for MGI (Fig. S9). We also compared the number and identity of the abundant OTUs found for each cutoff. The entire 16S rDNA dataset, the one resampled at 208 sequences, and the one resampled at 488 sequences, showed 17, 18, and 21 abundant OTUs (> 1%), respectively (19, 22, and 21 for the 16S rRNA sequences). Notably, the abundant OTUs were always the same in the different datasets.

Defining Abundant and Rare Phylotypes. OTUs were considered abundant when they comprised more than 1% of the sequences (11) and were present in more than one sample. In contrast, rare OTUs were defined as OTUs representing ≤0.2% of the sequences in a sample (present once in a sample of 488 sequences). This definition is well within the 0.1% to 1% range commonly considered (58), and is more strict than the 1% threshold used recently (4). OTUs were defined as always rare when they were rare in all the samples.

Representative sequences from all OTUs were compared with reference sequences from the entire SILVA database (56) using BlastN (59) to identify the percentage similarity between the queried sequences and their top hits. To assess the commonness of the sequences, the distribution of their percentage identity was plotted and fitted to normal distributions by using a maximum-likelihood method implemented in the mixture analysis of the PAST program (60). The method allowed us to define sequences as common (96–98% identity to database sequences) or uncommon (83% identity in average).

ACKNOWLEDGMENTS. We thank Cyrielle Tricoire, Eric Maria, and the captain and crew of the *Nereis II* for sample collection; and the people involved in the long-term series of hydrobiogeochemical data collected within the Service d'Observation en Milieu Littoral network (SOMLIT). This work was supported by a PhD fellowship from the French Ministère de l'Enseignement Supérieur et de la Recherche (to M.H.), a PhD fellowship from the French Conseil Régional d'Auvergne (to N.T.), and a CNRS Program Ecosphère Continentale et Côtière (EC2CO, 2010–2012). The work of P.E.G is supported by the Agence Nationale de la Recherche (ANR) project MICADO (ANR-11JSV7-003-01). J. Craig Venter Institute (JCVI) Global Ocean sampling, sequencing, and sequence analyses were funded by grants from the Beyster fund of the San Diego Foundation and the Life Technologies Foundation (to JCVI).

1. Alonso-Saez L, et al. (2007) Seasonality in bacterial diversity in north-west Mediterranean coastal waters: Assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiol Ecol* 60:98–112.
2. Mary I, et al. (2006) Seasonal dynamics of bacterioplankton community structure at a coastal station in the western English Channel. *Aquat Microb Ecol* 42: 119–126.
3. Campbell BJ, Kirchman DL (2012) Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. *ISME J* 7:210–220.
4. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL (2011) Activity of abundant and rare bacteria in a coastal ocean. *Proc Natl Acad Sci USA* 108:12776–12781.
5. Lennon JT, Jones SE (2011) Microbial seed banks: The ecological and evolutionary implications of dormancy. *Nat Rev Microbiol* 9:119–130.
6. Campbell BJ, Yu L, Straza TRA, Kirchman DL (2009) Temporal changes in bacterial rRNA and rRNA genes in Delaware (USA) coastal waters. *Aquat Microb Ecol* 57: 123–135.
7. Jones SE, Lennon JT (2010) Dormancy contributes to the maintenance of microbial diversity. *Proc Natl Acad Sci USA* 107:5881–5886.
8. Lami R, Ghiglione JF, Desdevises JF, West NJ, Lebaron P (2009) Annual patterns of presence and activity of marine bacteria monitored by 16S rDNA-16SrRNA fingerprints in the coastal NW Mediterranean Sea. *Aquat Microb Ecol* 54:199–210.
9. Gaidos E, Rusch A, Ilardo M (2011) Ribosomal tag pyrosequencing of DNA and RNA from benthic coral reef microbiota: Community spatial structure, rare members and nitrogen-cycling guilds. *Environ Microbiol* 13:1138–1152.
10. Sogin ML, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere" *Proc Natl Acad Sci USA* 103:12115–12120.
11. Pedros-Alio C (2006) Marine microbial diversity: Can it be determined? *Trends Microbiol* 14:257–263.
12. Galand PE, Casamayor EO, Kirchman DL, Lovejoy C (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc Natl Acad Sci USA* 106:22427–22432.
13. Pedros-Alio C (2012) The rare bacterial biosphere. *Annu Rev Mar Sci* 4:449–466.

14. Francis CA, Beman JM, Kuypers MM (2007) New processes and players in the nitrogen cycle: The microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J* 1:19–27.
15. Hugoni M, et al. (2013) Dynamics of ammonia-oxidizing *Archaea* and *Bacteria* in contrasted freshwater ecosystems. *Res Microbiol*, 10.1016/j.resmic.2013.01.004.
16. Karner MB, DeLong EF, Karl DM (2001) Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* 409:507–510.
17. Herndl GJ, et al. (2005) Contribution of *Archaea* to total prokaryotic production in the deep Atlantic Ocean. *Appl Environ Microbiol* 71:2303–2309.
18. Teira E, et al. (2008) Linkages between bacterioplankton community composition, heterotrophic carbon cycling and environmental conditions in a highly dynamic coastal ecosystem. *Environ Microbiol* 10:906–917.
19. Varela MM, van Aken HM, Sintes E, Herndl GJ (2008) Latitudinal trends of *Crenarchaeota* and *Bacteria* in the meso- and bathypelagic water masses of the Eastern North Atlantic. *Environ Microbiol* 10:110–124.
20. Massana R, DeLong EF, Pedros-Alio C (2000) A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. *Appl Environ Microbiol* 66:1777–1787.
21. Galand PE, Casamayor EO, Kirchman DL, Lovejoy C (2009) Unique archaeal assemblages in the Arctic Ocean unveiled by massively parallel tag sequencing. *ISME J* 3: 860–869.
22. Garcia-Martinez J, Rodriguez-Valera F (2000) Microdiversity of uncultured marine prokaryotes: The SAR11 cluster and the marine *Archaea* of Group I. *Environ Microbiol* 9:935–948.
23. Tully BJ, Nelson WC, Heidelberg JF (2012) Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. *Environ Microbiol* 14: 254–267.
24. Galand PE, Gutierrez-Provecho C, Massana R, Gasol J, Casamayor EO (2010) Interannual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea. *Limnol Oceanogr* 55:2117–2125.
25. Koeppel A, et al. (2008) Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci USA* 105:2504–2509.
26. Teske A, Sorensen KB (2008) Uncultured *archaea* in deep marine subsurface sediments: Have we caught them all? *ISME J* 2:3–18.
27. Bouvier T, del Giorgio PA (2007) Key role of selective viral-induced mortality in determining marine bacterial community composition. *Environ Microbiol* 9:287–297.
28. Jeraldo P, Chia N, Goldenfeld N (2011) On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys. *Environ Microbiol* 13: 3000–3009.
29. Ragan-Kelley B, et al. (2012) Collaborative cloud-enabled tools allow rapid, reproducible biological insights. *ISME J* 7:461–464.
30. Szabo G, et al. (2013) Reproducibility of Vibrionaceae population structure in coastal bacterioplankton. *ISME J* 7:509–519.
31. Coleman ML, Chisholm SW (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci USA* 107:18634–18639.
32. Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ (2010) The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. *Environ Microbiol* 12:490–500.
33. Brown MV, et al. (2012) Global biogeography of SAR11 marine bacteria. *Mol Syst Biol* 8:595.
34. Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB (2005) Ubiquity and diversity of ammonia-oxidizing *archaea* in water columns and sediments of the ocean. *Proc Natl Acad Sci USA* 102:14683–14688.
35. Wuchter C, et al. (2006) Archaeal nitrification in the ocean. *Proc Natl Acad Sci USA* 103:12317–12322.
36. Konneke M, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 437:543–546.
37. Lomas MW, Lipschultz F (2006) Forming the primary nitrite maximum: Nitrifiers or phytoplankton? *Limnol Oceanogr* 51:2453–2467.
38. Hallam SJ, et al. (2006) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci USA* 103:18296–18301.
39. Baker BJ, Lesniewski RA, Dick GJ (2012) Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. *ISME J* 6:2269–2279.
40. Konstantinidis KT, Braff J, Karl DM, DeLong EF (2009) Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl Environ Microbiol* 75:5345–5355.
41. Yakimov MM, et al. (2011) Contribution of crenarchaeal autotrophic ammonia oxidizers to the dark primary production in Tyrrhenian deep waters (Central Mediterranean Sea). *ISME J* 5:945–961.
42. Walker CB, et al. (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci USA* 107:8818–8823.
43. Auguet JC, Casamayor EO (2008) A hotspot for cold *crenarchaeota* in the neuston of high mountain lakes. *Environ Microbiol* 10:1080–1086.
44. Herfort L, et al. (2007) Variations in spatial and temporal distribution of *Archaea* in the North Sea in relation to environmental variables. *FEMS Microbiol Ecol* 62:242–257.
45. Robidart JC, et al. (2012) Seasonal Synechococcus and Thaumarchaeal population dynamics examined with high resolution with remote in situ instrumentation. *ISME J* 6:513–523.
46. Martens-Habbena W, Berube PM, Urakawa H, de la Torre JR, Stahl DA (2009) Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* 461:976–979.
47. Frigaard NU, Martinez A, Mincer TJ, DeLong EF (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and *Archaea*. *Nature* 439:847–850.
48. Iverson V, et al. (2012) Untangling genomes from metagenomes: Revealing an uncultured class of marine *Euryarchaeota*. *Science* 335:587–590.
49. Rusch DB, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5:e77.
50. Tanenbaum DM, et al. (2010) The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Stand Genomic Sci* 2:229–237.
51. Takai K, Horikoshi K (2000) Rapid detection and quantification of members of the archaeal community by quantitative PCR using fluorogenic probes. *Appl Environ Microbiol* 66:5066–5072.
52. Giongo A, et al. (2010) PANGEA: Pipeline for analysis of next generation amplicons. *ISME J* 4:852–861.
53. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194–2200.
54. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461.
55. Kim M, Morrison M, Yu Z (2011) Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods* 84: 81–87.
56. Pruesse E, et al. (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35: 7188–7196.
57. Taib N, Mangot JF, Domaizon I, Bronner G, Debroas D (2013) Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: New insights into the freshwater protist diversity. *PLoS ONE*, 10.137/journal.pone.0058950.
58. Fuhrman JA (2009) Microbial community structure and its functional implications. *Nature* 459:193–199.
59. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
60. Hammer Ø, Harper DAT, Ryan PD (2001) PAST: Paleontological Statistics Software Package for Education and Data Analysis. Available at http://folk.uio.no/ohammer/past/. Accessed May 20, 2009.

Supporting Information

Hugoni et al. 10.1073/pnas.1216863110

SI Materials and Methods

Nucleic Acid Extraction and Pyrosequencing. DNA and RNA were extracted by digesting cells directly in the Sterivex cartridge with lysis buffer (EDTA 40 mM; Tris 50 mM, pH 8.3, sucrose 0.75 M). Then heat/cold shocks treatments were performed, followed by the addition of lysozyme (40 mg·mL^{-1}) and incubation at 37 °C for at least 45 min. Then, proteinase K (20 mg·mL^{-1}) and SDS (0.2 g·mL^{-1}) were added, and the cartridges were incubated at 55 °C for 2 h. β-Mercaptoethanol was added to the cartridges content and nucleic acids extracted with the AllPrep DNA/RNA kit following the manufacturer's specifications (Qiagen). DNA and RNA yields were quantified by using a spectrophotometer (ND 1000; Nanodrop), and nucleic acid extracts were stored at −20 °C until analysis. RNA samples were tested for the presence of contaminant genomic DNA by PCR with archaeal 16S rDNA primers. Then, total RNA was reverse-transcribed with random primers by using the SuperScript III Reverse Transcriptase kit (Invitrogen) following the manufacturer's specifications. An identical set of reactions minus the reverse-transcriptase (i.e., no-RT reactions) were performed for each RNA extract; these reactions served as controls to examine the potential contributions of carryover genomic DNA on the PCR amplification of the cDNA. Nucleic acids were sent to Research and Testing Laboratory (Lubbock, TX) for amplification of the V3–V5 region of the 16S rRNA gene with universal archaeal primers Arch349F (CCC TAC GGG GTG CAS CAG) and Arch806R (GGA CTA CVS GGG TAT CTA AT) (1), followed by pyrosequencing on a Roche 454 GS-FLX system with titanium chemistry.

Pyrosequencing Data Analysis. ***Reference database.*** Sequences were compared with a dedicated database of reference sequences extracted from the SSURef 108 database from the SILVA project, which offers taxonomic information, quality assessment, and a curated alignment of SSU rRNA sequences (2). For the domain Archaea, the database includes 11,092 sequences with more than 900 bp, quality score >75%, and pintail value >50 according to the SILVA classification.

SILVA sequences together with annotated reference sequences from the Mediterranean Sea (3) were split into three monophyletic groups corresponding to the phyla Crenarchaeota, Thaumarchaeota, and Euryarchaeota, and a fourth group gathering sequences not affiliated to one of the three phyla. For each phyletic group, an outgroup containing one sequence from each of the other phyletic groups plus two distant sequences was added to the alignment to root the phyletic tree, and to specify the relatedness of early diverging sequences from the root of the group. Sequences from each phyletic group together with the outgroup sequences were retrieved from the SILVA alignment, and then trimmed to remove vertical gaps. A Hidden Markov Model (HMM) profile was built from each of the phyletic groups using HMMbuild from the HMMER package (4). A taxonomy file containing European Molecular Biology Laboratory taxonomy of each sequence from the reference database was also generated.

Sequence processing. First, a cleaning procedure was performed, whereby PANGEA functionalities (5) were used to remove short sequences (<200 bp), sequences with low quality score (≤27), sequences with at least one undetermined base, sequences with more than one mismatch with the forward primer, and to and trim tags and adaptors. Then, we used UCHIME (6) for the detection of chimera. From the 477,589 raw sequences, 414,579 were kept after cleaning and 407,053 after chimera checking.

Second, clean reads were then clustered with UCLUST (7) at 97% identity.

Third, the operational taxonomic units (OTUs) were compared against the reference database with USEARCH (7). Then, following the taxonomy of its best hit, each sequence was appended to a phyletic group, together with its five best hits. The query sequences were sorted according to their assignment.

Fourth, homologous reads were aligned with the referenced sequences from the corresponding profile using HMMalign (4). Next, a phylogenetic tree was build for each phyletic profile using FASTTREE2 (8) with the Jukes–Cantor + Cat model and a bootstrap threshold of 100.

Fifth, trees were parsed to generate files containing the taxonomy of the inserted sequences. The taxonomy assessment was inferred by lowest common ancestor. All sequences affiliated with Bacteria or that were unclassified were discarded from further analysis.

Finally, the pipeline produces a file containing the monophyletic clusters with their bootstrap values, a list of all affiliated experimental sequences, their nearest reference neighbor and their taxonomy.

The package used for this analysis (named PANAM) can be obtained from http://code.google.com/p/panam-phylogenetic-annotation/. It comprises the reference sequences database, the taxonomy file, and reference profile alignments.

Statistical Analyses. Canonical correspondence analysis (CCA) was performed in XL Stat 2010 (Addinsoft) to assess the relationships between taxonomic groups and environmental parameters. Reads for each taxonomic cluster considered were pooled for each sampling date. CCA was performed on 10 environmental factors (temperature, salinity, oxygen, Chl*a* content, pH, nitrite, nitrate, phosphate, silicate, and ammonium concentrations) with the taxonomic cluster abundance (inferred from read number) matrix of 16S rRNA dataset.

Kendall correlations were calculated to evaluate potential significant relationships between 16S rRNA and rDNA frequency for each OTU and time point (Fig. 1). Correlations were considered significant when $P < 0.05$.

SI Results

Environmental Parameters. Analyses of 3.5 y of environmental data from the coastal surface waters (3 m) of Banyuls-sur-Mer (France) indicated concentrations of inorganic nutrients and phytoplankton biomass characteristic of an oligotrophic environment during most of the year (Fig. S8). Temperature peaked in the late summer; on the contrary, oxygen concentrations were highest during winter. Some parameters displayed a consistent temporal pattern with highest abundance in winter, i.e., nitrate, silicate, or nitrite, and, to a lesser extent, Chl*a* (Fig. S8), illustrating the hydrodynamic features of the surface waters studied, particularly during winter and spring when precipitation and freshwater input from the land, and sediment resuspension after heavy storms, are recurrent events.

Comparison with Metagenomics Data. To assess possible bias, we compared pyrosequencing vs. metagenomics sequence counts from the J. Craig Venter Institute metagenomic analysis. The proportion of Euryarchaeota vs. Thaumarchaeota obtained by pyrosequencing for the September 28, 2010, sample was similar to the one obtained through the J. Craig Venter Institute metagenomic analysis (95% and 88% Euryarchaeota for pyrosequencing and metagenomics, respectively).

1. Takai K, Komatsu T, Inagaki F, Horikoshi K (2001) Distribution of *Archaea* in a black smoker chimney structure. *Appl Environ Microbiol* 67:3618–3629.
2. Pruesse E, et al. (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188–7196.
3. Galand PE, Gutierrez-Provecho C, Massana R, Gasol J, Casamayor EO (2010) Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea. *Limnol Oceanogr* 55:2117–2125.
4. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763.
5. Giongo A, et al. (2010) PANGEA: Pipeline for analysis of next generation amplicons. *ISME J* 4:852–861.
6. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194–2200.
7. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461.
8. Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490.

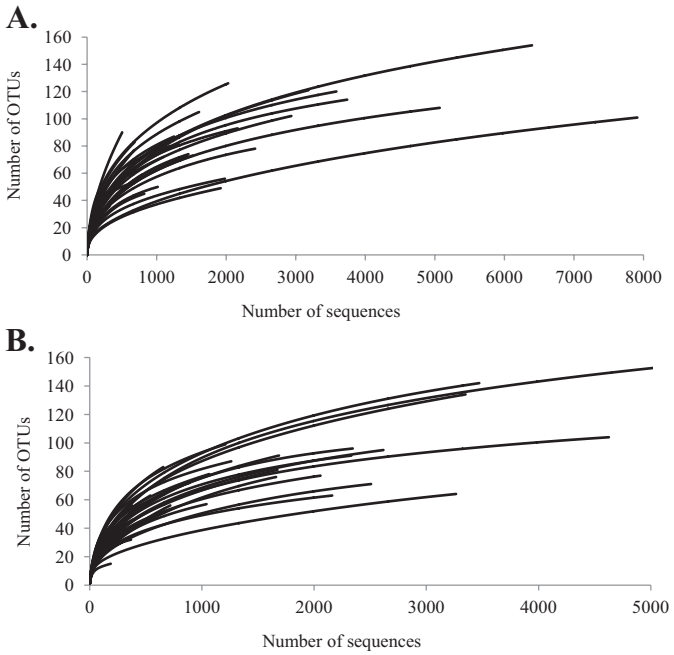


Fig. S1. Rarefaction curves for the 40 samples from the Banyuls-sur-Mer Bay coastal waters in the 16S rDNA dataset (*A*) and 16S rRNA dataset (*B*). The samples with too few sequences that were discarded from the analysis were far from saturation.

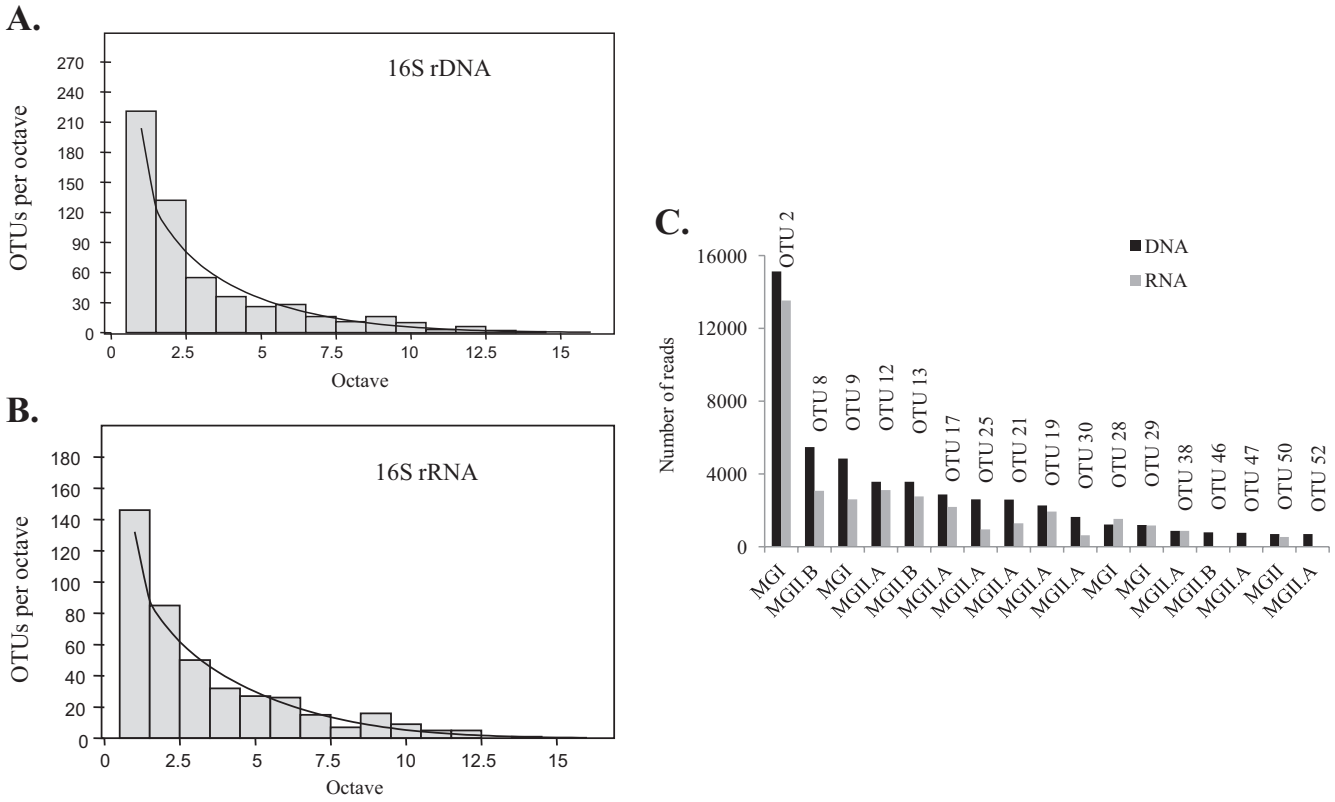


Fig. S2. Abundance distribution of archaeal OTUs obtained from the 40 samples from March 2008 to June 2011 at the Service d'Observation du Laboratoire Arago station in the 16S rDNA dataset (*A*) and the 16S rRNA dataset (*B*). The abundance models predicting the frequency of each abundance class are shown as lines. In both datasets, OTUs abundances were predicted by a log-series model. Octaves refer to power-of-two abundance classes. (*C*) The curve represents a detailed view of the respective abundant OTUs found in both 16S rDNA and 16S rRNA datasets.

Fig. S3. Phylogenetic tree representing the position of marine group (MG) I 16S rDNA sequences from the Banyuls-sur-Mer Microbial Observatory. Reference sequences retrieved from GenBank are in bold. Abundant OTUs are represented in red, and always-rare ones are in blue. Bootstrap values >70 are shown expressed as a percentage of 100 replicates. (Scale bar: 10% sequence divergence.)

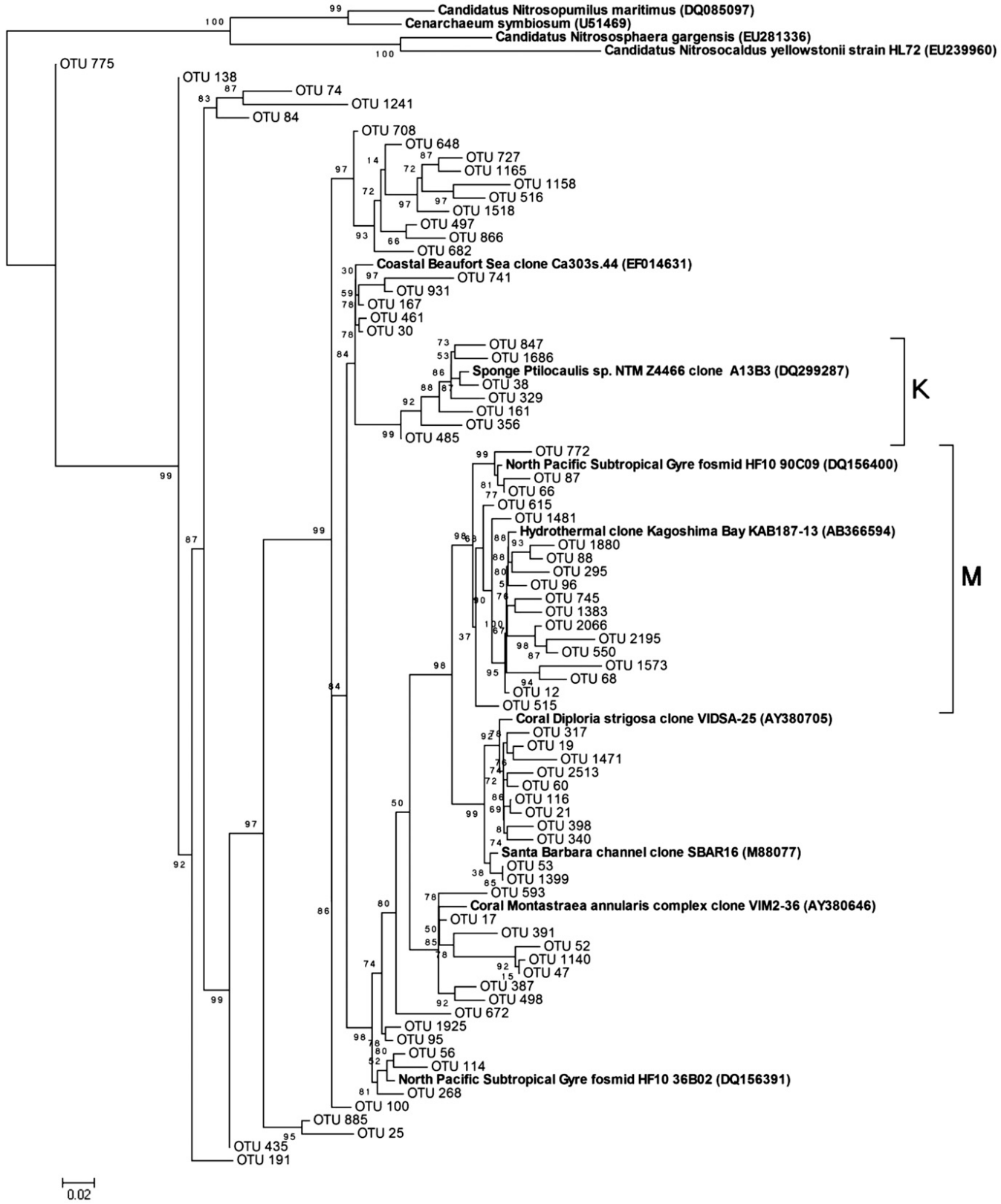


Fig. S4. Phylogenetic tree representing the position of MGII.A 16S rDNA sequences from the Banyuls-sur-Mer Microbial Observatory. Reference sequences retrieved from GenBank are in bold. Abundant OTUs are represented in red, and always-rare are ones in blue. Bootstrap values >70 are shown expressed as a percentage of 100 replicates. (Scale bar: 20% sequence divergence.)

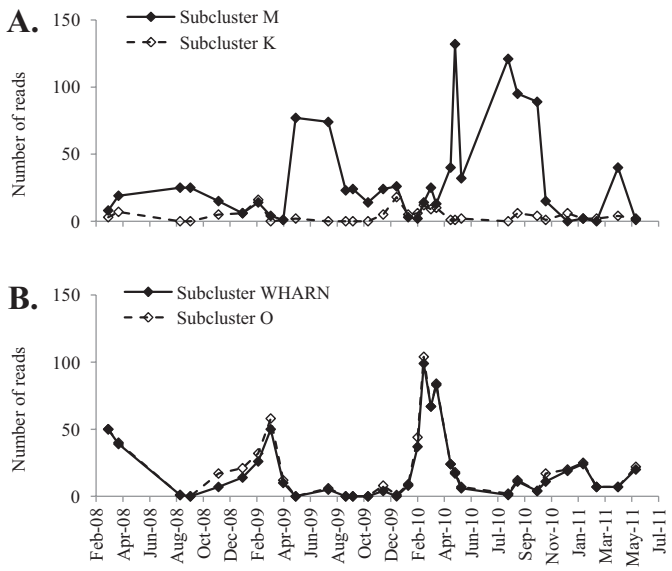


Fig. S5. Seasonal dynamics of active (16S rRNA dataset) MGII.A (*A*) and MGII.B subclusters (*B*). Winter months are shown in grey; summer months are shown in white.

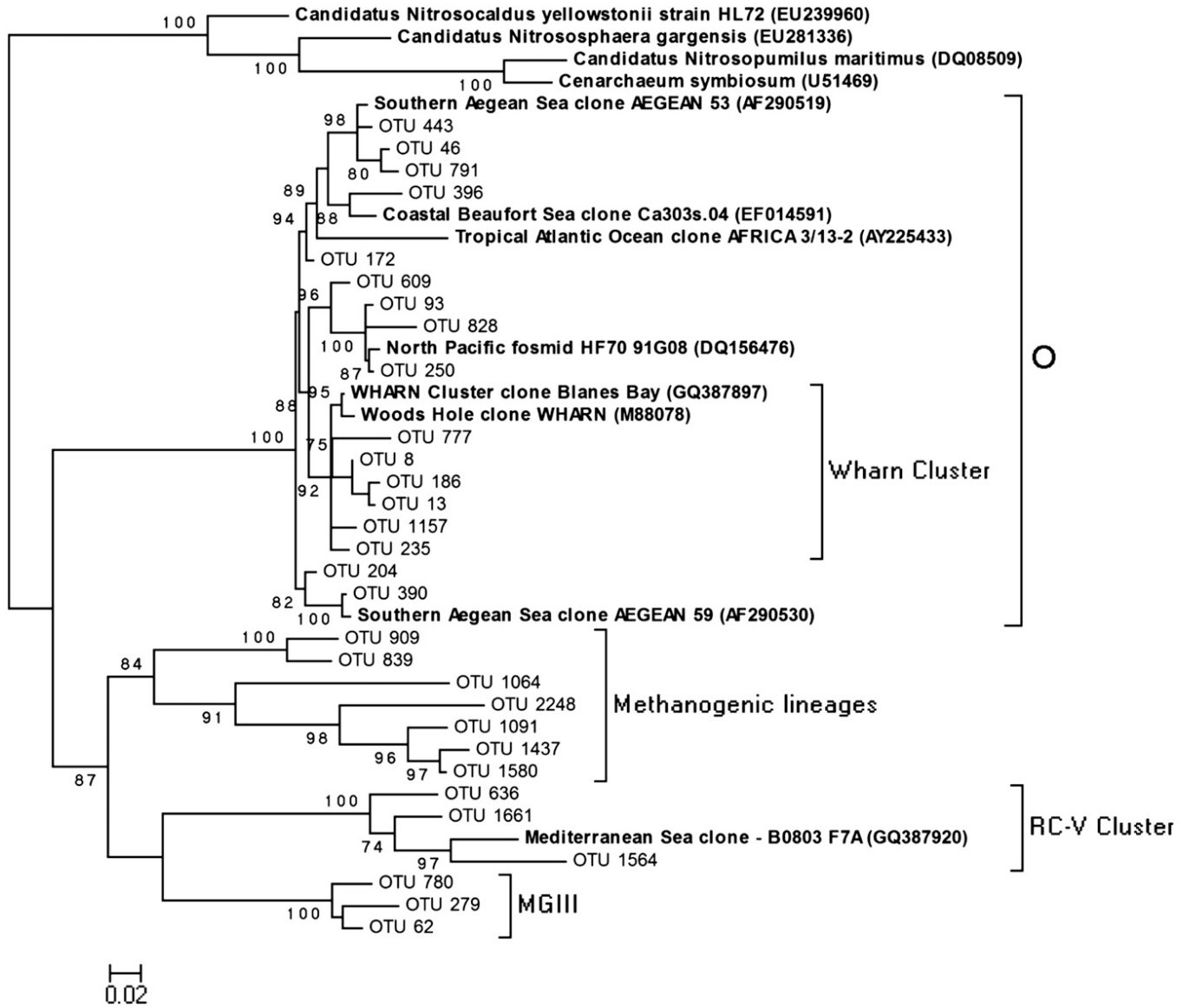


Fig. S6. Phylogenetic tree representing the position of MGII.B 16S rDNA sequences from the Banyuls-sur-Mer Microbial Observatory. Reference sequences retrieved from GenBank are in bold. Abundant OTUs are represented in red, and always-rare ones are in blue. Bootstrap values >70 are shown expressed as a percentage of 100 replicates. (Scale bar: 20% sequence divergence.)

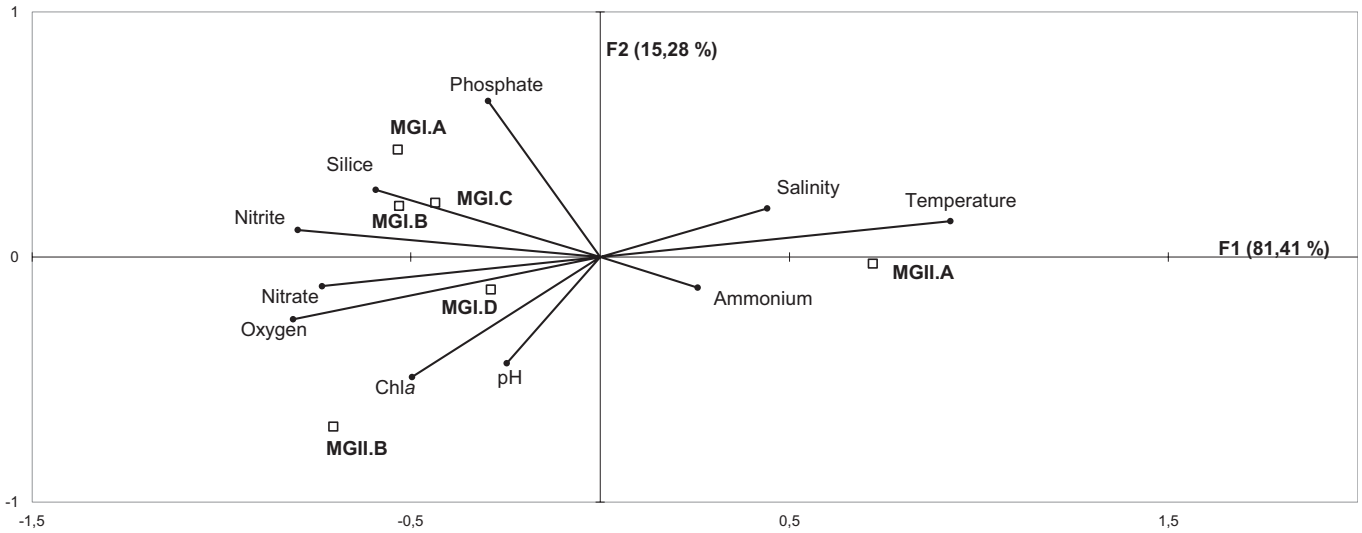


Fig. S7. Ordination diagram from CCA of 16S rRNA taxonomic clusters compared with environmental data.

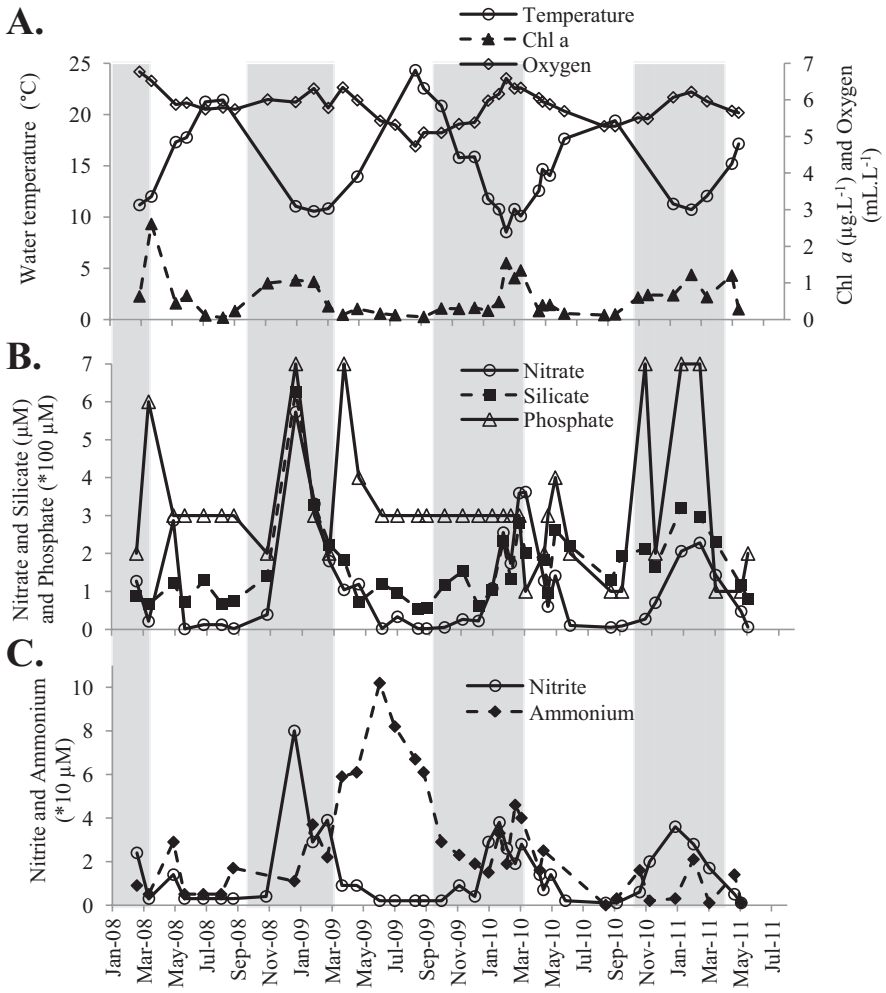


Fig. S8. Seasonal dynamics of environmental factors in the Bay of Banyuls-sur-Mer coastal waters. (*A*) Temperature, Chl*a*, and oxygen concentrations; (*B*) nitrate, phosphate, and silicate concentration; and (*C*) nitrite and ammonium concentrations. Winter months are shown in grey; summer months are shown in white.

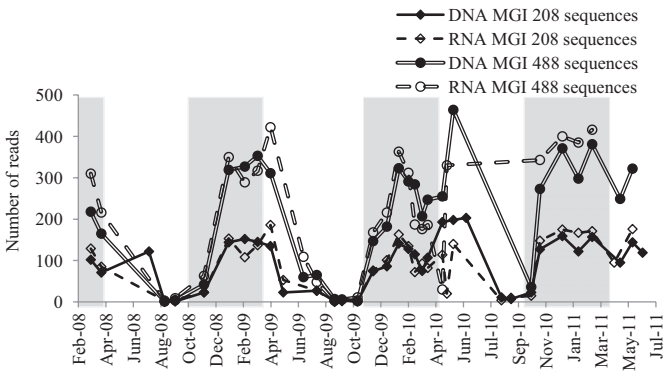


Fig. S9. Seasonal dynamics of MGI in both 16S rDNA and rRNA sequence datasets for the two normalization threshold 288 and 488 sequences. Winter months are shown in grey; summer months are shown in white.

Table S1. Raw read counts and QC reads obtained for each sample from surface seawater (3 m) collected monthly at SOLA station in the Bay of Banyuls-sur-Mer

	Raw sequences		QC sequences											
Date	16S rDNA	16S rRNA	16S rDNA	16S rRNA	Temperature, °C	Salinity, PSU	Oxygen, $mL \cdot L^{-1}$	pH	NH_4^+, µM	NO_3^-, µM	NO_2^-, µM	PO_4^{3-}, µM	$Si(OH_4)$, µM	Chla, $\mu g \cdot L^{-1}$
March 10, 2008	5,179	16,601	925	4,626	11.2	37.7	5.9	8.3	0.09	1.3	0.24	0.02	0.88	0.63
April 2, 2008	9,194	3,265	3,573	1,064	12	38.1	6.3	8.2	0.05	0.21	0.03	0.06	0.66	2.61
May 19, 2008	3,606	2,418	92	187	17.3	35.6	5.8	8.0	0.29	2.87	0.14	0.03	1.21	0.44
June 9, 2008	5,063	4,017	12	76	17.8	37.0	6.4	7.9	0.05	0.01	0.03	0.03	0.71	0.65
July 15, 2008	2,368	2,524	264	111	21.2	37.6	6	8.0	0.05	0.12	0.03	0.03	1.29	0.11
Aug. 18, 2008	6,267	4,295	1,979	2,161	21.2	37.9	5.4	8.2	0.05	0.12	0.03	0.03	0.67	0.05
Sep. 10, 2008	8,872	4,599	1,437	1,041	ND	ND	5.74	8.13	0.17	0.02	0.03	0.03	0.74	0.23
Nov. 12, 2008	3,523	4,085	658	639	ND	ND	6.01	8.23	ND	0.39	0.04	0.02	1.4	0.99
Jan. 5, 2009	10,104	7,075	3,583	2,333	11.0	36.4	5.3	7.9	0.11	5.73	0.8	0.07	6.27	1.07
Feb. 9, 2009	8,128	11,426	3,183	3,349	10.6	36.9	4.7	8.1	0.37	3.31	0.29	0.03	3.29	1.03
March 9, 2009	2,687	4,688	821	1,661	10.9	37.6	5.1	8.1	0.22	1.8	0.39	0.02	2.23	0.36
April 6, 2009	24,336	2,569	6,397	625	ND	ND	6.34	7.95	0.59	1.04	0.09	0.07	1.83	0.13
May 4, 2009	5,576	1,600	2,415	640	13.9	37.4	5.1	8	0.61	1.19	0.09	0.04	0.72	0.29
June 17, 2009	6,190	4,320	47	40	ND	ND	5.43	8.2	1.02	0.02	0.02	0.03	1.2	0.16
July 16, 2009	3,929	6,259	684	488	ND	ND	5.32	7.85	0.82	0.33	0.02	0.03	0.96	0.12
Aug. 24, 2009	11,151	5,576	7,910	3,264	24.3	37.8	5.4	7.9	0.67	0.02	0.02	0.03	0.53	ND
Sep. 9, 2009	3,715	2,636	1,015	368	22.6	38.0	5.4	8	0.61	0.02	0.02	0.03	0.56	0.07
Oct. 13, 2009	3,890	10,405	1,921	2,507	20.9	38.1	5.9	7.8	0.29	0.05	0.02	0.03	1.18	0.3
Nov. 16, 2009	4,923	21,839	1,251	3,471	15.8	38.2	6.2	8.2	0.23	0.26	0.09	0.03	1.53	0.29
Dec. 16, 2009	4,501	8,548	1,193	1,261	15.9	38	6.6	8.3	0.19	0.22	0.04	0.03	0.61	0.32
Jan. 11, 2010	4,455	2,593	1,609	655	11.8	38.0	6.3	8.2	0.15	1.08	0.29	0.03	1.03	0.24
Feb. 1, 2010	9,785	16,820	3,741	5,065	10.8	37.6	6.3	8.3	0.33	2.55	0.38	0.03	2.33	0.48
Feb. 15, 2010	4,140	7,208	1,519	2,057	8.5	37.5	6.0	8.3	0.19	1.74	0.26	0.03	1.32	1.54
March 3, 2010	18,607	5,760	5,071	1,672	10.8	36.3	5.9	8.3	0.46	3.59	0.19	0.03	2.81	1.13
March 15, 2010	8,379	3,035	2,936	894	10.1	36.9	5.9	8.3	0.4	3.62	0.28	0.01	2	1.34
April 16, 2010	4,674	4,443	327	541	12.6	37.5	5.7	8.2	0.16	1.28	0.14	0.02	1.84	0.23
April 26, 2010	4,238	3,100	88	719	14.7	37.3	5.3	8.1	0.25	0.6	0.07	0.03	0.96	0.39
May 10, 2010	5,567	6,206	792	1,687	14.1	37.3	5.3	8.3	ND	1.41	0.14	0.04	2.62	0.4
June 7, 2010	6,589	5,339	351	104	17.7	37.4	5.5	8.3	ND	0.1	0.02	0.02	2.2	0.16
Aug. 23, 2010	2,195	3,025	227	247	ND	ND	5.28	8.12	0	0.05	0.01	0.01	1.29	0.12
Sep. 13, 2010	3,219	7,419	235	386	19.4	38.00	5.5	8.3	0.03	0.09	0.01	0.01	1.93	0.14
Oct. 27, 2010	3,258	1,067	1,099	291	ND	ND	5.51	8.26	0.16	0.27	0.06	0.07	2.13	0.6
Nov. 15, 2010	5,726	4,740	2,027	1,210	ND	ND	5.48	8.29	0.02	0.7	0.2	0.02	1.64	0.67
Jan. 3, 2011	3,467	7,532	1,457	2,618	11.3	37.6	6.1	8.1	0.03	2.06	0.36	0.07	3.2	0.66
Feb. 7, 2011	3,937	6,955	1,310	2,344	10.7	37.4	6.2	8.2	0.21	2.28	0.28	0.07	2.96	1.22
March 9, 2011	7,508	3,038	2,166	823	12.0	38.1	5.9	8.3	0.01	1.43	0.17	0.01	2.3	0.61
April 26, 2011	2,647	3,354	127	378	15.2	37.1	5.7	8.2	0.14	0.47	0.05	0.01	1.16	1.2
May 9, 2011	6,405	2,712	505	122	17.1	37.5	5.6	8.2	0.01	0.06	0.01	0.02	0.81	0.28
June 5, 2011	3,767	3,495	678	413	18.07	38.1	5.44	8.11	0.01	0.03	0.02	0.01	0.7	0.21
June 27, 2011	6,835	2,403	208	43	19.8	37.8	5.32	8.4	0.01	0.08	0.01	0.01	0.47	0.11

Environmental parameters (temperature, salinity, oxygen, pH, ammonium, nitrate, nitrite, phosphate, silicate, and Chla) associated to each point are presented. ND, not determined; QC, quality-checked; SOLA, Service d'Observation du Laboratoire Arago.

Article 4

Short-term dynamics of diversity patterns : evidence of continual reassembly within lacustrine small eukaryotes

environmental microbiology

Environmental Microbiology (2012) doi:10.1111/1462-2920.12065

Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes

Jean-François Mangot,[1,2,3,4] Isabelle Domaizon,[1,2†] Najwa Taib,[3,4†] Nemr Marouni,[3,4] Emilie Duffaud,[3,4] Gisèle Bronner[3,4] and Didier Debroas[3,4*]

[1]*INRA, UMR 42 CARRTEL, Thonon les bains, France.*

[2]*Université de Savoie, UMR 42 CARRTEL, Le Bourget du Lac, France.*

[3]*Clermont Université, Université Blaise Pascal, Laboratoire 'Microorganismes: Génome et Environnement', BP 10448, F-63000 Clermont-Ferrand, France.*

[4]*CNRS, UMR 6023, LMGE, F-63177 Aubière, France.*

Summary

The short-term variation in the community structure of freshwater small eukaryotes (0.2–5 μm) was investigated in a mesotrophic lake every 2–3 days over one summer by coupling three molecular methods: 454 amplicon pyrosequencing, qPCR and TSA-FISH. The pyrosequencing approach unveiled a much more extensive small-eukaryotic diversity (991 OTUs) than has been described previously. The vast majority of the diversity described was represented by rare OTUs (≤ 0.01% of reads) belonging primarily to *Cryptomycota*, *Dikarya* and photosynthetic organisms, which were never detected as abundant in any of the samples. The small eukaryote community was characterized by a continual and important reassembly. These rearrangements involved the 20 'core taxa' (≥ 1% of reads), and, were essentially due to a handful of OTUs that were detected in intermediate abundance (0.01–1% of reads) and sporadically in dominant taxa. Putative bacterivorous (*Ciliophora* and *Cercozoa*) as well as parasitic and saprotrophic taxa (*Perkinsozoa* and *Cryptomycota*) were involved in these changes of diversity. A putative infection of microalgae by a lacustrine perkinsozoan was also reported for the first time in this study. Open questions regarding both the patterns that govern the rapid small eukaryote reassemblies and the possible biogeography of these organisms arise from this study.

Received 23 December, 2011; revised 3 October, 2012; accepted 22 November, 2012. *For correspondence. E-mail didier.debroas@univ-bpclermont.fr; Tel. (+33) 4 73 40 78 37; Fax (+33) 4 73 40 76 70. †These authors contributed equally to this work.

Introduction

Small eukaryotes (i.e. cells ≤ 2, 3 or 5 μm in size fraction according to the studies) play a major role in biogeochemical cycles, especially the global carbon cycle in marine environments (Liu *et al.*, 2009) and are likely the most abundant eukaryotes on Earth (Zhao *et al.*, 2011). Over the last decade, thanks to the emergence of molecular techniques in microbial ecology, the diversity and abundance of these small eukaryotes have been investigated both in marine (e.g. López-García *et al.*, 2001; Not *et al.*, 2008) and more recently in freshwater systems (e.g. Lefranc *et al.*, 2005; Lepère *et al.*, 2008). From these approaches, the recurrent presence of parasites and quantitative importance of pigmented groups have been revealed (Mangot *et al.*, 2009; Lepère *et al.*, 2010). Furthermore, these works have highlighted an unexpected diversity with the existence of new environmental clades in lakes belonging, for instance, to *Perkinsozoa* (clades 1 and 2) or *LKM11* recently classified among the *Cryptomycota* (Jones *et al.*, 2011).

However, these approaches, which have been performed at specific times and locations, may only reflect a partial picture of the eukaryotic picoplankton community. Because of their rapid growth (cells divide up to once per day or more), picoplanktonic populations (essentially the photosynthetic component) may respond rapidly to environmental fluctuations (e.g. Jacquet *et al.*, 1998; Vaulot and Marie, 1999), being dependent on biotic interactions (Reckermann and Veldhuis, 1997) or direct and indirect effects of viral attacks (Cottrell and Suttle, 1995). Chambouvet and colleagues (2008) recently revealed dinoflagellate–parasitoid successions in a natural estuary in correlation with the rapid development of their host populations, which may suggest rapid shifts in *in situ* parasite dynamics. Such short-term variations could therefore occur in freshwater ecosystems where the importance of putative parasitoids has been underscored. Therefore, studies performed over short-time scales appear to be critical for obtaining a better understanding

of the factors that control and regulate these eukaryotic populations. Several recent investigations performed at an intermediate time scale (1- or 2-week to 1-month intervals) on the entire protistan community have already revealed the regularity and rapidity with which the eukaryotic assemblage restructures itself to yield unique combinations of dominant taxa (Vigil *et al*., 2009; Nolte *et al*., 2010). According to Caron and Countway (2009), these rapid shifts among protistan community may be the result of rare taxa that become dominant with changing environmental conditions. Massively parallel sequencing techniques now provide an in-depth analysis of microbial diversity and consequently offer opportunity to investigate the putative ecological role of this rare biosphere (Sogin *et al*., 2006; Galand *et al*., 2009).

For this purpose, the short-term variation in the diversity and abundance of freshwater small eukaryotes has been investigated in a mesotrophic lake every 2–3 days during one summer by combining the 454 amplicon pyrosequencing, qPCR and TSA-FISH approaches. We hypothesize that microbial interactions (i.e. predator–prey and host–parasite associations, competition for resources . . .) could induce rapid shifts among small-planktonic protists. Massively parallel sequencing constitutes a powerful method, however, according to various authors, the amplicon pyrosequencing approach still provides only a limited comparative analysis across samples (Amend *et al*., 2010; Gifford *et al*., 2010; Zhou *et al*., 2011). Therefore, in this study, and based on the methodology proposed by Gifford and colleagues (2010), a normalized amplicon pyrosequencing approach was employed.

Results

Physicochemical and meteorological characteristics

The average water temperature during the sampling period was 17.3°C [coefficient of variation (CV) = 4.5%], and a constant concentration in dissolved oxygen (11 mg l^{-1}, CV = 7.9%) was measured between 0 and 20 m (Table 1). The mean concentrations of PO_4-P, NH_4-N and NO_3-N were 0.04 μM (CV = 32.5%), 0.62 μM (CV = 44.8%) and 4.48 μM (CV = 17.8%) respectively. Strong solar irradiance (182.7 KJ m^{-2} day^{-1} on average) and high-wind regimes (mean = 14.6 m s^{-1}) were recorded throughout the sampling period, with occasional days of cloudy weather (i.e. the days 24 and 47 of sampling; Fig. S1). Finally, despite 4-weak rainfall events, the study period was characterized by important dry episodes (Fig. S1). This study period was globally characterized by significant variations of environmental variables in comparison with the seasonal changes known in Lake Geneva (long-term monitoring).

Table 1. Main meteorological and physico-chemical characteristics (mean and CV) of Lake Geneva over the study period (from 17 July to 18 September 2009).

Parameters	Mean	CV (%)
Meteorology[a]		
Solar radiation (KJ m^{-2} day^{-1})	182.7	34.6
Wind speed (m s^{-1})	14.6	31.9
Physicochemistry[b]		
Dissolved oxygen (mg l^{-1})	11	7.9
Water temperature (°C)	17.3	4.5
NO_3-N (μM)	0.61	17.8
NH_4-N (μM)	4.49	44.8
PO_4-P (μM)	0.04	32.5

a. Daily means.
b. Means integrating the 0–20 m layer.

Quantification of the total small eukaryotes by combining different molecular approaches

Read data were significantly correlated with the fluctuations of 18S rDNA copies ml^{-1} estimated by qPCR ($r = 0.64$; $P = 0.001$; Fig. 1); however, without normalization, no significant correlations were recorded ($r = 0.16$; $P = 0.48$). Depending on which molecular tool was applied, different perceptions of the small eukaryote dynamics were obtained (Fig. 1). The quantification of the small eukaryote community, which was performed by qPCR and normalized pyrosequencing, revealed strong temporal changes, with five common and sudden increases in the number of 18S rDNA copies recorded over the study period (days 1, 9, 20, 31, 45 and 52; Fig. 1A). Thus, by qPCR, we observed a decrease from 253 740 to 64 copies ml^{-1} in a few days (days 45 and 47), which represents a 3905-fold decrease in the number of 18S rDNA copies. Finally, the small eukaryotes targeted by TSA-FISH, which varied from 4673 cells ml^{-1} (day 5) to 20 675 cells ml^{-1} (day 14), were characterized by a lower fluctuation in abundance (a maximum of a 3.5-fold change between two consecutive dates; Fig. 1B) compared with the other descriptors. According to the TSA-FISH counting and the small subunit (SSU) 18S rRNA gene copy numbers of eukaryotes as estimated by qPCR or normalized reads, we estimated a mean of 5.3 and 6.3 copies of SSU rRNA genes by cell.

Short-term variations in the richness and diversity of the small eukaryote community

Over the study period, 991 operational taxonomic units (OTUs) were determined. The small eukaryote richness showed variation marked by serial decrease–increase events in a short time (CV = 34.3%; Fig. 2). The Shannon index fluctuated between 0.75 (day 5) and 4.58 (day 24), and it showed, in general, the same serial decrease–increase events that were observed for the richness ($r = 0.71$; $P = 0.0002$).

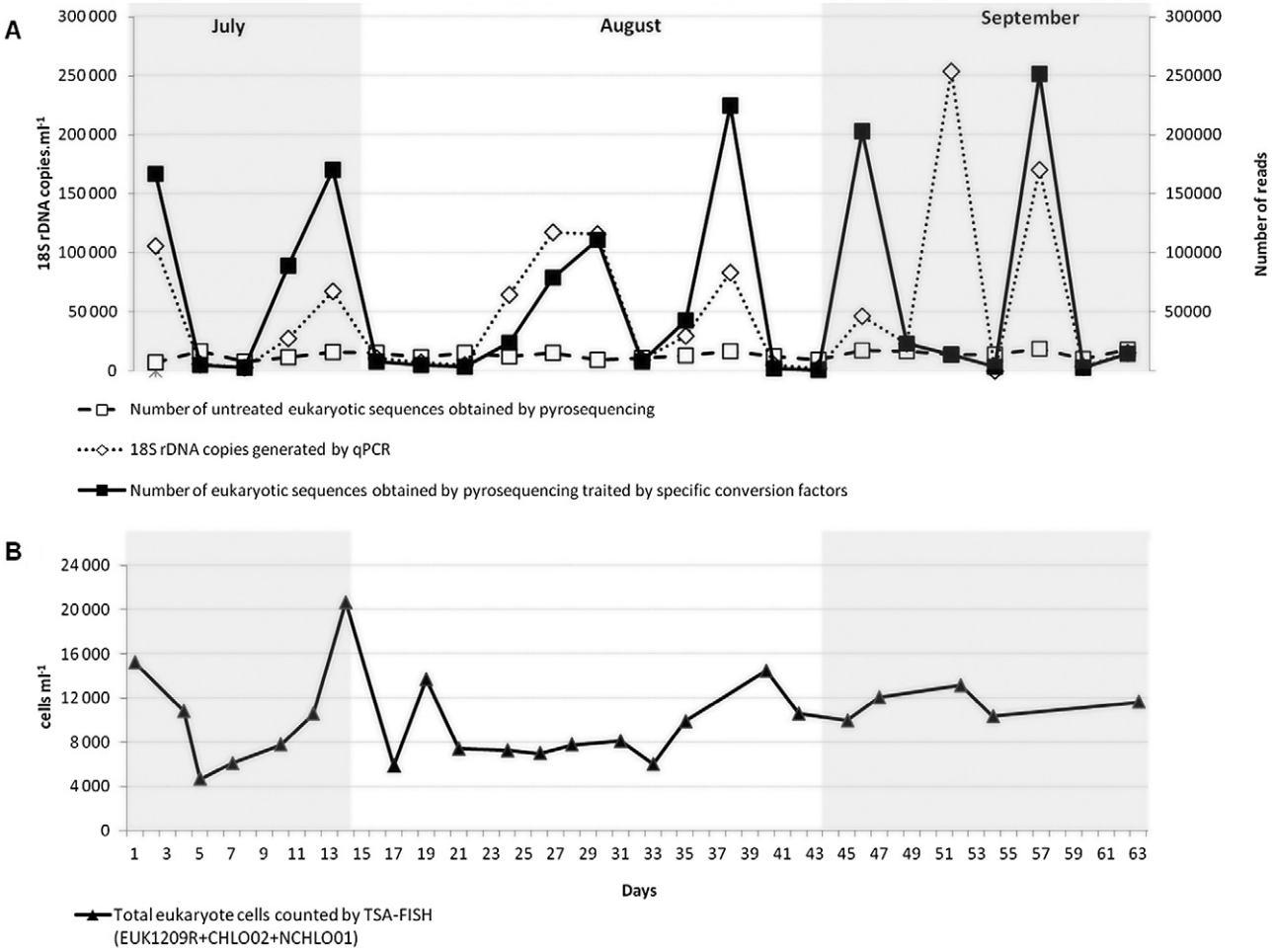


Fig. 1. Short-term dynamics (every 2 or 3 days) of the small eukaryote community in Lake Geneva (0–20 m) obtained by different molecular approaches (qPCR, pyrosequencing and TSA-FISH) over the 63 days of experimentation (from 17 July to 18 September 2009).
A. Small eukaryotes represented by the short-term variation in the total 18S rDNA genes estimated by qPCR (copies ml^{-1}) and by pyrosequencing before (number of reads) and after the application of sample-specific conversion factors that were calculated by means of a quantitative internal standard (number of reads).
B. Small eukaryote dynamics estimated by the variations of cells targeted by TSA-FISH by the mix of probes (Euk1209R + CHLO02 + NCHLO01; cells ml^{-1}).

The dissimilarity calculations carried out at the community scale showed important temporal changes in the small eukaryote composition over the time with a mean of 65.8% dissimilarity between two consecutive sampling dates (CV = 24.5%; Fig. 3). Regarding the abundant taxa, important temporal changes in their composition were recorded over the study period (mean = 30.4% dissimilarity; CV = 52.8%). However, a relative stability (less than 20% dissimilarity on average) was reported between days 7 and 26. In the same manner, the dissimilarity calculations performed for the 'intermediate taxa' populations fluctuated from 13.4% to 64.4% dissimilarity between two consecutive sampling days (CV = 36.5%). Finally, the small eukaryote rare biosphere composition revealed low dissimilarity indexes between two consecutive dates (mean = 1% dissimilarity).

Taxonomic composition of the small eukaryote assemblage

Almost 27% of the OTUs were affiliated with the alveolates and more precisely the *Ciliophora* and *Perkinsozoa* taxa (Fig. 4; Table 2). The number of reads associated with these phyla represented 32.2% and 14.9% of the sequences respectively. By using specific probes, the *Perkinsozoa* clades 1 and 2 were found to represent 1.1% of this assemblage (Table 3). After the alveolates, the small eukaryote diversity consisted of fungi and stramenopiles, which represented 29.2% and 20.5% of the OTUs respectively. Heterokonts sequences were exclusively composed of bacilliariophycean and chrysophycean taxa. Among the fungi, sequences affiliated with the *Cryptomycota*, *Dikarya* (*Ascomycota* and *Basidiomycota*)

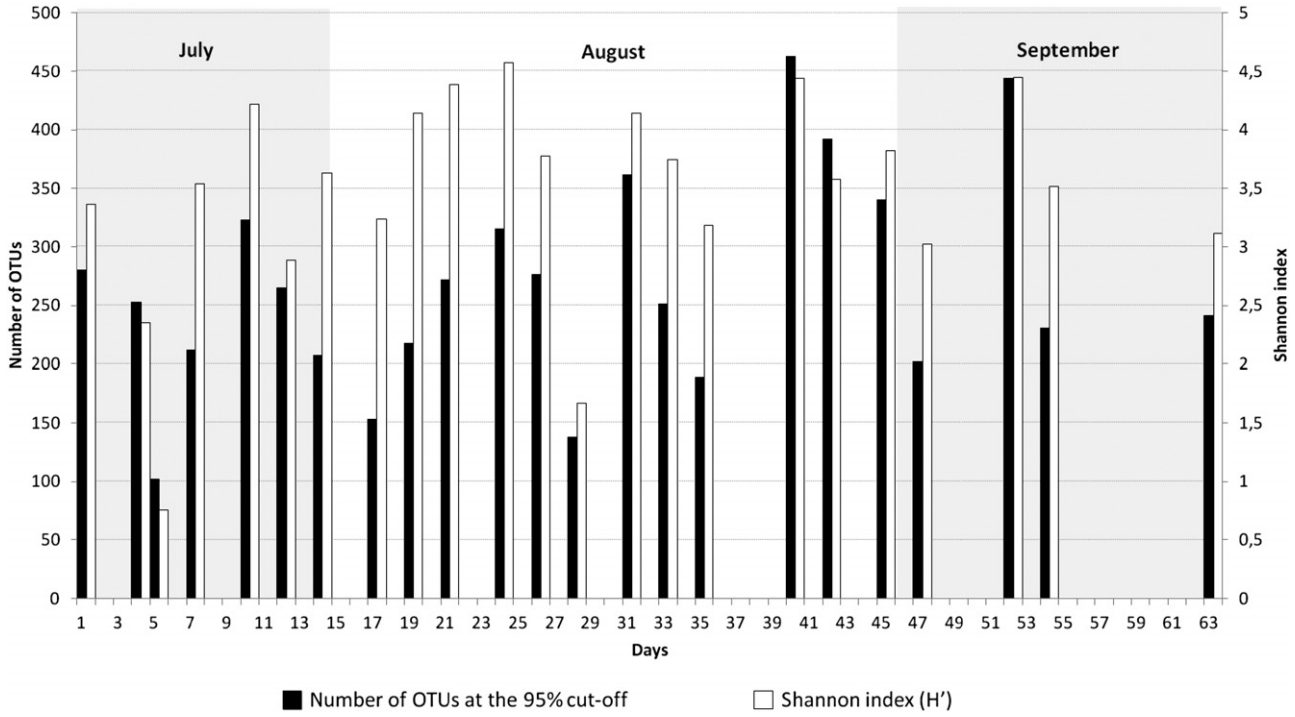


Fig. 2. Short-term variations in the richness (number of OTUs) and the diversity estimator (Shannon index) of the small eukaryote community over the 63 days of experimentation.

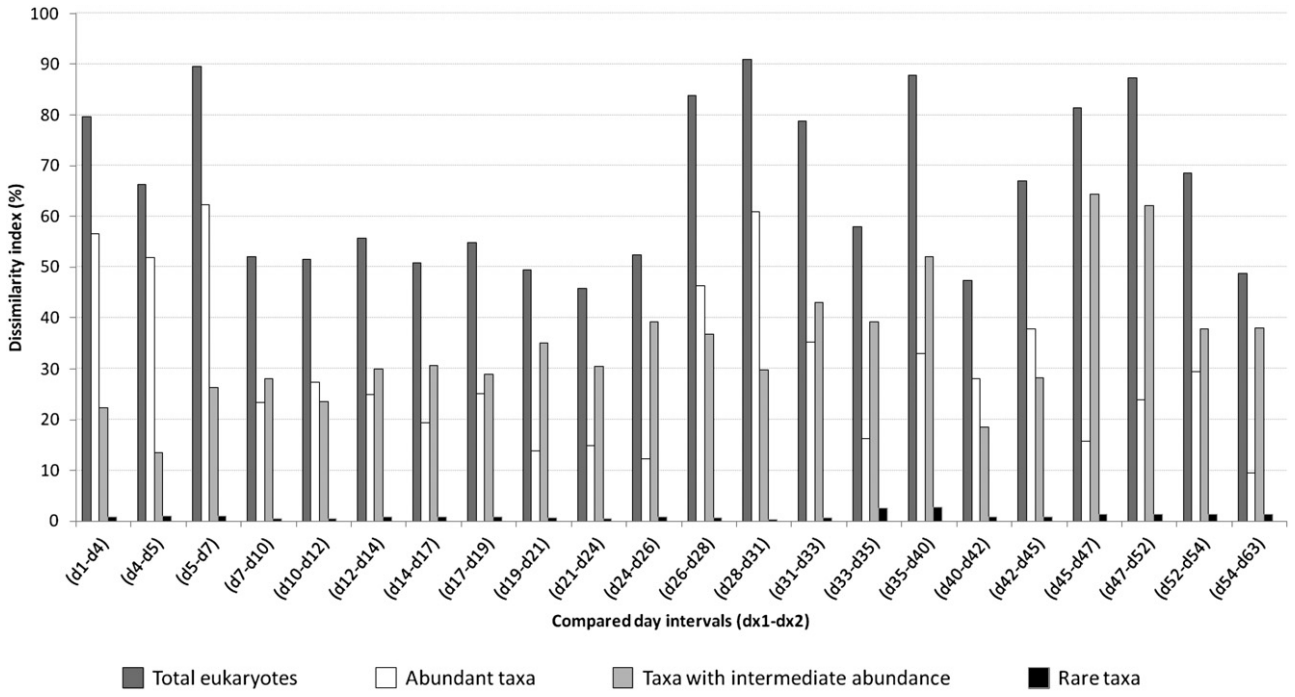


Fig. 3. Dissimilarity indexes of the total small eukaryote community, abundant (≥ 1%), intermediate (between 1 and 0.01%) and rare taxa (≤ 0.01% of total reads), over the 63 days of experimentation. Indexes are calculated for each consecutive sampling day over the 23 sampling dates [noted (dx1–dx2), with dx1 and dx2 as two consecutive sampling dates].

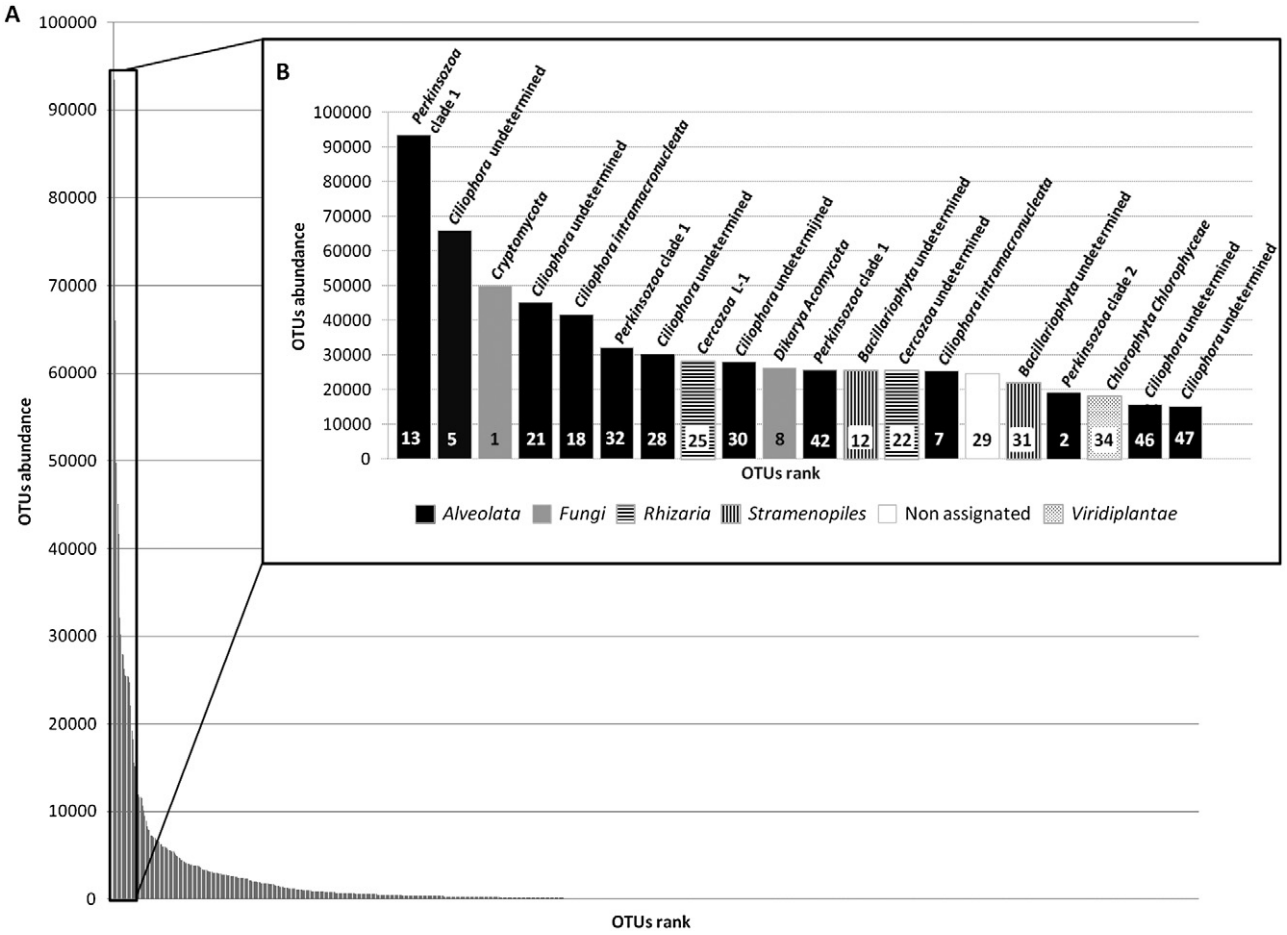


Fig. 4. Rank abundance curve of the small eukaryote operational taxonomic units (OTUs) for the combined dataset of 208 319 partial 18S rRNA gene sequences from the 23 samples obtained from 17 July to 18 September 2009 in Lake Geneva.
A. The curve represents the respective abundance of all of the 991 OTUs defined and (B) of the 20 'core taxa', which accounted for more than 1% of the total reads. The identification number of the 20 abundant OTUs and their taxonomic affiliation are also mentioned in and on the bars respectively.

and *Chytridiomycota* groups were primarily retrieved (Table 2). The fungi represented on average 16.5% of the reads and 13.1% of the abundance as determined by TSA-FISH, among which respectively 6.9% and 9.7% are only constituted by the *Cryptomycota* (Table 3). Other small eukaryotes (slightly more than 20% of both the sequences and the OTUs) were composed of cercozoan sequences (8.7% of the reads) and OTUs affiliated with the *Chlorophyceae* (2.5% of the OTUs). Therefore, the mean abundance of *Cercozoa*, determined by the FISH method (6.6%), was similar to the proportion of reads, whereas the two molecular methods used were divergent for *Chlorophyceae*. Finally, approximately 3.4% of the entire small eukaryote community (up to 3.5% of the OTUs) was still non-assigned. Among the 991 OTUs, between 0 (using USEARCH) and 7% (using the furthest neighbour algorithm implemented in MOTHUR) was detected in other lacustrine ecosystems by traditional Sanger method (data from public databases with a 95% cut-off). This analysis highlighted the presence of freshwater clades (Table 2) previously defined in Lefranc and colleagues (2005) and Lepère and colleagues (2008), but also the presence of numerous OTUs non-detected previously. Finally, almost 57% of the small eukaryote community quantified by TSA-FISH (Table 3) was represented by eukaryotic cells not detected by our set of probes (listed in Table S1) and belonged to other groups, for instance, to *Choanoflagellida*, *Amoebozoa*, *Cryptophyta*, *Haptophyceae*, *Stramenopiles* . . . , as suggested by sequencing results (Table 2).

Twenty OTUs were designated as 'core taxa' over all of the 23 samples (Fig. 4B). These OTUs each contributed at least 1% (and together represented up to 48%) of the reads. Sixty per cent of these 20 abundant taxa belonged to the two alveolates groups *Ciliophora* and *Perkinsozoa*. The rest of the abundant groups over the study period were mainly represented by opisthokonts (*Cryptomycota* and *Ascomycota*), cercozoan (two OTUs, among which

Table 2. Taxonomic composition and diversity (relative of normalized reads and number of OTUs) of the small eukaryote community characterized in Lake Geneva (0–20 m) from 17 July to 18 September 2009 by pyrosequencing.

				Number of OTUs at the 95% cut-off			
Taxonomic affiliation			Reads	Total	Abundant	Intermediate	Rare (singleton)
Alveolata	*Apicomplexa*	unclassified *Apicom.*	0.8%	4		3	1
		Aconoidasida	0.0%	6			6 (3)
		Coccidia	0.0%	1			1
	Ciliophora	unclassified *Ciliop.*	22.2%	76	6	42	28 (11)
		Alveolate L-1*	2.2%	29		19	10 (3)
		Alveolate L-2*	0.5%	3		2	1
		Colpodea	0.3%	5		3	2 (1)
		Intramacronucleata	6.6%	56	2	22	32 (9)
		Scuticociliadida	0.5%	1		1	
	Dinophyceae		3.4%	52		25	27 (13)
	Perkinsea		0.2%	9		6	3
		Perkinsea clade 1*	13.2%	20	3	13	4 (2)
		Perkinsea clade 2*	1.5%	5	1	2	2 (1)
Amoebozoa	*Centramoebida*	*Acanthamoebidae*	0.1%	3		1	2 (2)
	Tubulinea	*Euamoebida*	0.0%	5			4 (2)
Choanoflagellida			1.5%	28		15	13 (3)
Cryptophyta	*Cryptomonadaceae*		0.4%	5		4	1 (1)
	Cryptophyta	nucleomorph	0.0%	4			4 (1)
		Cryptophyta L-3*	0.1%	3		3	0
		Cryptophyta L-4*	0.6%	4		3	1
Euglenozoa			0.0%	1			1
Fungi	*Fungi*	unclassified *Fungi*	1.7%	51		34	17 (5)
	Chytridiomycota		1.0%	18		10	8 (3)
	Cryptomycota[a]		6.9%	95	1	30	64 (18)
	Dikarya	unclassified *Dikarya*	0.1%	23		8	15 (7)
		Ascomycota	4.6%	53	1	20	32 (11)
		Basidiomycota	1.9%	41		15	26 (10)
	Nowakowskiella clade		0.0%	2		1	1
	Tremellales et rel.		0.3%	5		2	3 (2)
	Zygomycota_1 et rel.		0.0%	1			1 (1)
Haptophyceae			0.9%	5		2	3 (1)
Rhizaria	*Cercozoa*	unclassified *Cercoz.*	3.9%	21	1	17	3 (1)
		Cercozoa nuclear	0.7%	14		8	6 (2)
		Cercozoa L-1*	3.4%	29	1	9	19 (10)
		Cercozoa L-2*	0.6%	7		5	2 (1)
Stramenopiles	*Stramenopiles*	unclassified *Stram.*	1.0%	33		25	8 (4)
	Bacillariophyta		4.4%	32	2	14	16 (7)
	Bicosoecida		1.6%	52		25	27 (18)
	Chrysophyceae	unclassified *Chryso.*	3.6%	46		33	13 (4)
		Chrysophyceae L-1*	0.1%	6		3	3 (3)
	Dictyochophyceae		0.3%	14		8	6 (3)
	Eustigmatophyceae		0.1%	4		2	2
	Labyrinthulida		0.3%	1		1	
	Oomycetes		0.3%	15		8	7 (6)
Viridiplantae	*Chlorophyta*	unclassified *Chloro.*	2.3%	36		19	17 (7)
		Chlorophyceae	2.5%	25	1	10	14 (3)
		Prasinophyceae	0.0%	3		2	1
		Trebouxiophyceae	0.0%	2		1	1
		Ulvophyceae	0.0%	1			1 (1)
Unclassified			3.4%	35	1	9	25 (14)
Total				991	20	486	485 (195)

a. Sequences affiliated to *LKM11* and *Rozella* clade with the EMBL taxonomy have been grouped within the newly defined term of *Cryptomycota*. The three taxonomic ranks constituted arbitrary rank defined in the sequence database from EMBL. The presence of freshwater clades is marked by an asterisk (*).

were the freshwater clade *Cercozoa* L-1) and chlorophycean OTUs. Furthermore, one non-assigned taxon and two OTUs belonging to bacilliariophycean (a stramenopiles group) were retrieved among these 'core taxa'. In contrast, 485 OTUs (approximately 48.9% of the total OTUs and approximately 0.7% of the reads) contributed to less than 0.01% of the pyrosequencing dataset. Among these rare taxa, 195 OTUs were constituted by singletons

Table 3. Small eukaryote abundances (mean and CV) targeted by TSA-FISH from 17 July to 18 September 2009 in Lake Geneva (0–20 m).

Targeted group	Mean (cells ml^{-1})	CV (%)
Total small eukaryotes (Euk1209R + CHLO02 + NCHLO01)	9859	36.6
Cercozoa (CERC_02)	649	46.6
Chlorophyceae (CHLO02)	2245	79.4
Fungi (MY1574)[a]	327	76.4
Cryptomycota (LKM11_01 + LKM11_02)	961	50.4
Perkinsozoa clade 1 (PERKIN_01)	54	57.9
Perkinsozoa clade 2 (PERKIN_02)	55	44.6

a. Except *Cryptomycota*.

(approximately 19.7% of the total OTUs; Table 2). Between these two entities (abundant and rare taxa), a third class of less abundant taxa was defined as 'intermediate taxa' and represented almost half of the OTUs.

Globally, only eight OTUs (less than 1% of total number of OTUs) were observed in all 23 of the samples, while 28.2% (279 of 991) of the OTUs were observed only at a single date. Furthermore, only 19.6% of the described taxa were common to more than half of the samples. We note that some freshwater clades were never recorded in the abundant class (≥ 1%), such as *Alveolate* L-1 and L-2.

The most abundant OTU detected is affiliated to perkinsozoan

Perkinsozoan sequences dominated the pyrosequencing dataset with four OTUs among the 20 'core taxa' defined (Fig. 3), and this group always appeared among the dominant groups throughout the study (Fig. S2). The highest abundances of *Perkinsozoa* clades 1 and 2 (detected by TSA-FISH method) were recorded on the first day of sampling (17 July) with 215 cells ml^{-1}. However, there was a discrepancy between the sequencing and FISH results. The pyrosequencing dataset indicated that clades 1 and 2 represented 13.2% and 1.5%, respectively, whereas the abundances of each clade estimated by TSA-FISH were approximately 0.6%. The alignments of the sequences affiliated with *Perkinsozoa* indicated a poor recovery of the probe specific of clade 1 in this ecosystem: 1.4% of the group diversity without mismatch and 76.2% with two mismatches. In contrast, the second probe allowed the capturing 96.2% of the diversity among clade 2 without mismatch. Intracellular non-flagellated perkinsozoan cells (PERKIN_02 probe) were observed inside a colonial *Volvocales*-like chlorophycean species (Fig. 5). On day

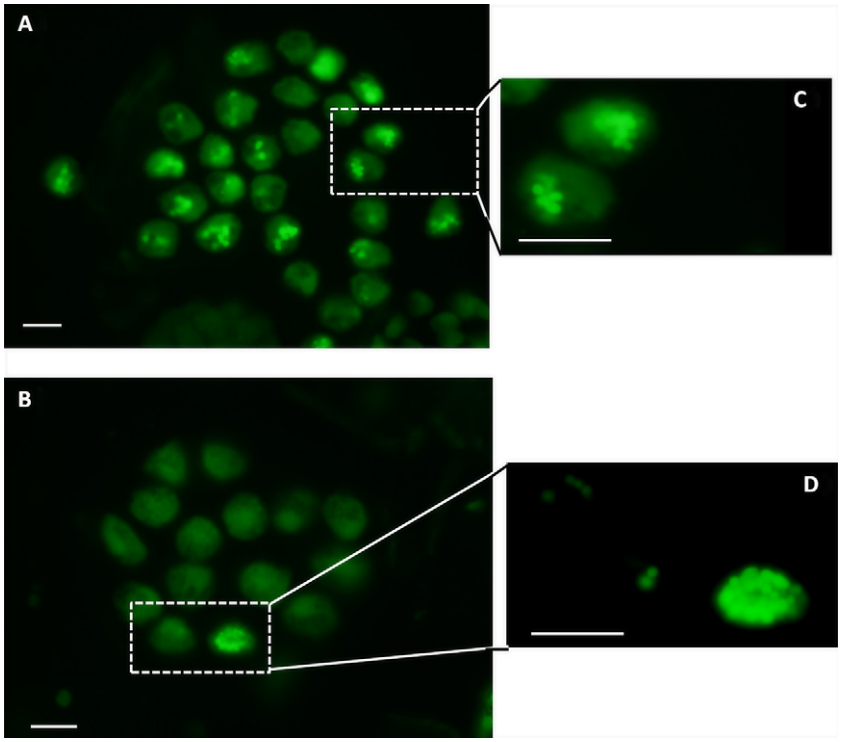


Fig. 5. Observations under blue light excitation of *Perkinsozoa* clade 2 cells targeted by TSA-FISH inside a chlorophycean *Volvocales*-like species (A and B, at the colony scale; C and D, at the cell scale). Scale bars, 10 μm.

45, approximately 13% of the *Volvocales*-like colonies have revealed an intracellular staining with the TSA-FISH probe specific of the *Perkinsozoa* clade 2. However, this apparently low infection rate at the *Volvocales*-like population scale was counterbalanced by a higher infection rate within the colony. Indeed, 72% of cells that formed the colony may appear to be infected by perkinsozoan cells (two to more than 34 intracellular cells per chlorophycean cells).

Discussion

A growing number of investigations have focused on the composition and structure of the protistan community within marine and, more rarely, freshwater systems, including some analyses performed by the 454 method (Amaral-Zettler *et al*., 2009; Nolte *et al*., 2010; Monchy *et al*., 2011). In this study, a characterization of the short-term variation of the lacustrine surface small eukaryote community ($\leq$ 5 μm) was provided, and constitutes, to our knowledge, the first investigation of these small protists and fungi in freshwater systems through a short-time-scale approach.

Contribution of molecular methods to the study of protist dynamics

For previous metagenomics or amplicon pyrosequencing approaches (Sogin *et al*., 2006; Huber *et al*., 2007; Urich *et al*., 2008; Monchy *et al*., 2011), the power of comparative analysis was still limited by the variability in sample-sequencing depth and the potential systematic biases inherent to the sampling processes, amplification and sequencing (Amend *et al*., 2010; Gifford *et al*., 2010). To perform meaningful comparisons across our different samples, based on the methodology proposed by Gifford and colleagues (2010), an internal rDNA standard was used, allowing for an estimation of the fraction of the microbial amplicons that were captured in the sequences library and, therefore, allowing for the quantification of the 18S rDNA present at each sampling date. Furthermore, the sequencing error as determined by the internal standard suggests an accurate clustering threshold of 95%. The justifications for the threshold values used in previous studies on the eukaryote community (Sogin *et al*., 2006; Not *et al*., 2009; Cheung *et al*., 2010; Monchy *et al*., 2011) have often been omitted, and the implication has been that they approximate species-level distinctions (Caron *et al*., 2009). Thus, the use of an internal DNA standard allows to (i) reliably define OTUs and therefore avoid bias in richness estimation (Caron *et al*., 2009) and (ii) quantify the microbial community dynamics.

Depending on which molecular tool was applied, different perceptions of the small eukaryote dynamics were obtained. Thus, the lower temporal variation recorded by the FISH method compared with 18S abundances could be due to the copy number of SSU rRNA genes, which varies widely among eukaryotes and seems to be correlated with cell length (Zhu *et al*., 2005). The 18S rDNA copy number by cell determined in this study is close to the value proposed by Cheung and colleagues (2010) and Zhu and colleagues (2005) in coastal systems (less than 10 copies per picoeukaryote). These low-estimated values of the average copy number of small eukaryotes' SSU rRNA genes allow, to a certain extent, a less biased analysis of the pyrosequencing datasets through rank-abundance reasoning. However, the dominance of alveolates and the low levels of pigmented organisms in the sequencing dataset could be explained by the well-described high copy number of the SSU rRNA genes in alveolates which is contrasted by a presumably low copy number in most flagellates (Zhu *et al*., 2005; Medinger *et al*., 2010) or, by certain PCR biases (Zhu *et al*., 2005; Not *et al*., 2009). In addition, the efficiency of FISH probes to target the whole cells within each group is questionable. For example, numerous sequences that were obtained by the high-throughput method in this study were not efficiently targeted by the PERKIN_01 probe which was designed previously with sequences from the Sanger method obtained from other lakes. Furthermore, difference in the ribosomes content of targeted cells might exist leading to underestimate the quantitative importance of some groups by TSA-FISH as already shown on *Escherichia coli* cultures (Hoshino *et al*., 2008). Finally, the observed inconsistencies between microscopical (FISH) and molecular surveys (qPCR, pyrosequencing) may be explained by the specificity of each method. Indeed, qPCR and pyrosequencing methods quantify DNA presumably of living, and to some extent dead cells, while TSA-FISH probes target only living and active cells (Kock and Schippers, 2008). Variations in qPCR, pyrosequencing and TSA-FISH quantification may be explained both by DNA preservation and/or degradation (Kock and Schippers, 2008), the variability in rRNA operons among the species (Lozupone and Klein, 2002; Lefèvre *et al*., 2010), the variability in the cell lysis's efficiency (Medinger *et al*., 2010), and by biogeography patterns for small- and microeukaryotes (Bik *et al*., 2012).

New insights into the structure of small eukaryote community

In our 454 amplicon library data, we highlighted a much greater number of OTUs than those found by the traditional Sanger method (e.g. Richards *et al*., 2005; Tarbe *et al*., 2011). Thus, a high-throughput pyrosequencing approach for 18S rRNA gene hypervariable regions,

which was primarily initiated to increase the sampling effort of rDNA gene sequences, allowed us to unveil an extensive eukaryote diversity in lakes (Monchy *et al*., 2011) that has been missed by the low-throughput approach (Stoeck *et al*., 2009). Despite the difference of sequencing depth between these two molecular methods, a similar phylogenetic composition at different phylogenetic levels, from the main phyla to the typical freshwater clades, was retrieved. However, in our study, few OTUs (most 7%) described over the study period were already present in public databases, which illustrated the importance of undetected diversity by previous diversity surveys, and/or possibly a biogeography for these microorganisms (no sequences were obtained from Lake Geneva before this study). In addition, we were able to explore the rare biosphere among the smallest lacustrine protists, which has already been demonstrated for prokaryote (Sogin *et al*., 2006; Galand *et al*., 2009) and eukaryote communities (Dawson and Hagen, 2009; Stoeck *et al*., 2009). The intermediary class (0.01–1%) was composed of taxa that belonged notably to the groups of *Ciliophora*, *Perkinsozoa*, *Cryptomycota* (former *LKM11* and *Rozella* groups) and *Cercozoa*, which at times briefly became dominant during the study period. The 'core taxa' (only 20 OTUs) were therefore mainly composed of putative parasites and saprotrophs (*Perkinsozoa*, *Cryptomycota*, *Dikarya*) and potential bacterivorous groups (*Ciliophora* and *Cercozoa*). A large fraction of these sequences belonged to potential parasites of freshwater phytoplankton and to ciliates, and they have already been found by other authors in the small size fraction (0.2–5 μm) by the cloning-sequencing approach (Lefèvre *et al*., 2007; 2008; Lepère *et al*., 2008). Among these dominant taxa, numerous sequences are affiliated to alveolates taxa (60% of the abundant OTUs) which are known, as stated above, to be overestimated in molecular datasets due to their high 18S copy number. The presence of ciliates could also derive from cell debris or extracellular DNA from larger cells (Not *et al*., 2009) and their presence among the 'core taxa' may rather be artificial than biological. However, small bactivorous and algivorous ciliates (< 20 μm) may reached 53% of overall ciliates numbers in marine systems (Mironova *et al*., 2012) and, could represent an important component of the small eukaryote assemblage that has been previously neglected.

A large part of the diversity described over the study belonged to the rare biosphere. This community is constituted by a vast majority of taxa that were always rare (99.8%), i.e. never detected as abundant in any of the samples analysed. This finding is in good agreement with that of Galand and colleagues (2009) and provides further information about the taxonomic composition and the putative ecological role of the rare biosphere in this system. Indeed, the rare biosphere was mainly composed of fungal sequences (*Cryptomycota* and *Dikarya*) and, to a lesser extent, photosynthetic organisms (*Chlorophyta*, *Bacilliarophyta* and *Dynophyceae*). The reality and the importance of this rare biosphere, and notably of unique sequences, were recently discussed, and numerous works suggested an overestimation of this community resulting from methodological artefacts, such as errors of sequencing and/or potential errors of taxonomic assignment due to an excessive division of OTUs by a non-adapted threshold (Quince *et al*., 2009; Reeder and Knight, 2009; Huse *et al*., 2010). To avoid such overestimation of the community diversity, most papers based on next-generation sequencing approach tend to remove these unique sequences to the total dataset (e.g. Medinger *et al*., 2010; Nolte *et al*., 2010). In our work, by the definition of an appropriate cut-off of 95% thanks to an internal standard DNA, we freed ourselves from the limitation of the diversity estimation inherent to the used technique. Also, we decided to preserve singletons which represent 19.7% of the total richness and do not affect our perception of the present small-eukaryotic composition. Indeed, these unique sequences are affiliated to unarguable groups constituted by other sequences which are present in much higher numbers in our dataset (Table 2). In recent years, numerous questions have emerged regarding the source (dead cells or molecular debris rather than viable cells), composition and persistence (active or non-active resting stages) of the assemblage composed by rare taxa (Caron and Countway, 2009; Galand *et al*., 2009; Jones and Lennon, 2010; Campbell *et al*., 2011). Dormancy is a bet-hedging strategy used by a variety of organisms to overcome unfavourable environmental conditions, and it plays a more important role in shaping bacterial communities than eukaryotic microbial communities (Jones and Lennon, 2010). According to our data, we cannot comment on the activity or inactivity of cells in the rare biosphere and the role of dormancy (low activity) among this assemblage. However, regarding the taxonomic composition of the rare taxa community, the presence of both active and dormant cells may be hypothesized. Indeed, the important presence in the euphotic zone of pigmented organisms and putative parasites and saprotrophs (notably *Cryptomycota*), which are present in a zoosporic form in this size fraction (Jones *et al*., 2011), suggests the presence of activity among these groups. However, some forms of dormancy among the picoplanktonic size fraction may be suggested by the important presence of fungal sequences (belonging to *Dikarya*) which are known to disseminate in the form of conidia (Jobard *et al*., 2010), and present low activity (Jones and Lennon, 2010).

Short-term reassembly of small eukaryotes did not involve members of the rare biosphere

The results obtained by TSA-FISH revealed up to a 4.4-fold variation in the small eukaryote density over a short-term period. Relatively similar amplitudes of variation in the total eukaryotic cells targeted by TSA-FISH (1692–10 782 cells ml^{-1}) were previously observed over a 1-year study (Mangot *et al*., 2009). These short-term changes were also associated with changes in the small eukaryote richness and diversity characterized over the study period. Indeed, a continual community reassembly was manifested by the high dissimilarity index encountered over the study period (Fig. 3), in accordance with previous observations of protistan assemblages over short temporal or spatial scales by fingerprint methods (Vigil *et al*., 2009). As hypothesized by Nolte and colleagues (2010), the small eukaryote community does not solely undergo quantitative fluctuations within a stable set of taxa that are present throughout the entire study period. From this study, we assume that the instability of the small eukaryote community that was observed in the short-time series should be considered carefully to avoid biasing the interpretation regarding the dynamics of these microorganisms, especially when limited sampling is performed to characterize the microbial community. Caron and Countway (2009) hypothesized that the taxa from the rare biosphere were involved in the community reassembly. In our study, we showed that the continual reassembly among the dominant taxa is essentially due to important exchanges between 'intermediate' and abundant taxa. The rare taxa were therefore not involved at this time scale, but they could possibly be involved for more drastic changes in environmental conditions.

Short-term diversity rearrangements under influence of abiotic and biotic factors

Rapid growth allows the microeukaryotes to respond rapidly to even minor environmental fluctuations (Countway *et al*., 2005; Caron and Countway, 2009). Nutrient concentrations and climatological factors recorded in Lake Geneva over the study period were characterized by non-negligible variations compared with the rest of the year (Alpine Lake Observatory – data not shown). In this study, we could not explain the variation in abundance or diversity for the different taxonomic groups, but it is clear that these small eukaryotes are strongly linked to the other components of the planktonic food web through their nutritional strategy. For example, algal exudates control bacterial abundances (Larsson and Hagström, 1979) and indirectly control bacterivorous organisms such as *Cercozoa*. Some abiotic parameters and biotic factors, such as predation or host availability for parasitic groups, were recognized to influence the growth rate and diversity of algae and their parasites (i.e. *Perkinsozoa* and fungi; Rasconi *et al*., 2011).

In this study, we focused on the microbial interactions involving the *Perkinsozoa* that dominated the small eukaryote community composition. This study constitutes therefore the first observations of putative infections of phytoplanktonic cells by a perkinsozoan species in a lacustrine system. *Perkinsozoa* clade 2, newly defined by Lepère and colleagues (2008) and Mangot and colleagues (2011), is structured around the marine species *Parvilucifera infectans*, which is known to infect various microalgal cells of which numerous belong to different dinoflagellate genera (Park *et al*., 2004). One study has previously reported the presence of an algal-parasitic *Perkinsozoa*, *Rastrimonas subtilis* gen. et sp. nov. (Brugerolle, 2002; Brugerolle, 2003), which is a parasite of cryptophytes in river systems. In this study, we suspected a parasitic activity on colonial *Chlorophyceae*, and, although its dynamic was not investigated, we hypothesize that the parasitic activity may have played a role in the short-term rearrangements observed over the study period.

The understanding of the short- and long-term dynamics of the small eukaryote diversity cannot be considered without taking into account the other planktonic compartments. We hypothesize, as has also been suggested recently by Lindström and Langenheder (2011) that the assembly mechanisms that shape the small protist community structure could vary according to functional groups; it is certainly a hypothesis to be tested for a better understanding of the patterns of protistan diversity.

Experimental procedures

Study site and sampling

This study was conducted in Lake Geneva (46°27′N, 06°32′E), a mesotrophic lake described in detail in Anneville and colleagues (2002). Water samples were taken at a permanent station in the 0–20 m layer using a sampling bell to carry out the integrated sampling throughout the water column (Pelletier and Orand, 1978). The sampling was performed in the summer every 2 or 3 days at the same hour (07:00 h) from 17 July to 18 September 2009. During the entire sampling period, daily climatological data such as the precipitation, wind speed and solar radiation were obtained at the meteorological observatory station of the INRA's Thonon station (France). Primary nutrients (PO_4-P, NO_3-N, NH_4-N) were measured in the freshwater samples by the chemical laboratory of the Thonon hydrobiological station according to French normalized (AFNOR) protocols (http://thononin8.win3.hebergement.com/). The temperature (°C) and dissolved oxygen concentration (mg l^{-1}) were measured using a Seabird submersible multiparametric probe with a CTD SBE 19 Seacat profiler. The water samples (at least 100 ml) were

sequentially filtered onto 100 μm and 5 μm membranes, and the filtrate (< 5 μm) was concentrated on 0.2 μm membranes and preserved for DNA analyses or counting by TSA-FISH according to our previous studies (Lepère *et al.*, 2008; 2010; Mangot *et al.*, 2009).

Furthermore, in order to examine microbial interactions (e.g. potential association of perkinsozoan with phytoplanctonic groups) sequential concentration of water samples (20 l) were performed by using 160 μm, 50 μm, 20 μm filter membranes, and 0.2 μm CellTrap™ filter cartridges (Mem-Teq, Orrell, UK). Each concentrate corresponded to a planktonic size fraction and was preserved for TSA-FISH observations as previously described.

Small eukaryote quantification by TSA-FISH

The small eukaryote community and five phylogenetic groups of freshwater eukaryotes, listed in Table S1, were quantified. TSA-FISH technique was performed exactly as described by Lepère and colleagues (2008) (see Supporting information S1). Hybridized cells were examined under blue light (480/535 nm) with a Zeiss Axiovert 200 M inverted and epifluorescence microscope (Carl Zeiss, Jena, Germany) equipped with an HBO 100 W Hg vapour lamp at × 100 magnification. For each sample, at least 50 randomly chosen microscopic fields were analysed and counted manually (on average, a minimum of 50 cells were counted).

DNA extraction and quantification of 18S rDNA copies

All of the samples were extracted following the protocol described previously by Lefranc and colleagues (2005). Quantification of the 18S rDNA copies was performed in duplicate by quantitative PCR by using sub-type 4 *Blastocystis* (Accession Number: FJ666885) plasmids as a standard. For this purpose, 18S rDNA standards were generated by the amplification of DNA extracted from *Blastocystis* using the universal eukaryotic primers EukA (5′-AACCTGGTTG ATCCTGCCAGT-3′) and EukB (5′-TGATCCTTCTGCAGG TTCACCTAC-3′) according to Díez and colleagues (2001) and cloned using a TOPO-TA cloning kit (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. Quantification was performed using the eukaryotic primers Ek-NSF573 (5′-CGCGGTAATTCCAGCTCCA-3′) and Ek-NSR951 (5′-TTGGYRAATGCTTTCGC-3′), which were chosen in our tag pyrosequencing approach (see below). The qPCR reaction volumes were 25 μl and were composed of 12.5 μl of MESA GREEN qPCR MasterMix Plus for SYBR Assays® (1×, Eurogentec, San Diego, CA, USA), 0.1 μM of each primer (Ek-NSF573 and Ek-NSR951), 2 mg ml^{-1} of bovine serum albumin (Fermentas, Sankt Leon-Rot, Germany) and 1 μl of template (DNA extract or plasmid). For the environmental samples, the DNA was diluted to obtain 30 ng μl^{-1}. The quantitative PCR cycling conditions were 10 min at 94°C (initial denaturation) followed by 45 cycles at 94°C (30 s), 60°C (30 s) and 72°C (45 s) using a Mastercycler®ep realplex real-time PCR system. Data were collected during the annealing phase. A mean standard curve efficiency of 118% was obtained, with a CV of 0.08%.

Pyrosequencing

To estimate the fraction of the microbial amplicons that were captured in the sequence library (i.e. the sample-sequencing depth), an internal 18S rDNA standard (Gifford *et al.*, 2010) was added before the amplification of each sample by a set of barcoded primers. This standard was *Blastocystis* (Accession Number: FJ666885) which is absent in freshwater samples. This 18S rDNA included in a plasmid (TOPO vector) was added to represent around 1% of 18S rDNA copies determined by qPCR in each sample. For instance 33.1 copies ml^{-1} of the *Blastocystis* plasmids have been added before the PCR amplification in day 54 according the copies number of 18S RNA genes determined at this date (3432 copies ml^{-1}). The proportions of our marker per sample were variable and represented on average 1.65% of 18S rDNA copies (CV = 185%).

The V4–V5 hypervariable region of eukaryotic 18S rDNA was amplified with Ek-NSF573 and Ek-NSR951 as previously described. This primer couple was selected by an *in silico* approach because it allowed for the best recovery of richness from the freshwater small eukaryote sequences available in public databases. To discriminate each sample, a 10 bp multiplex tag was coupled with adaptor A. The amplification mix (30 μl) contained 30 ng of genomic DNA (sample + internal standard), 200 μM of deoxynucleoside triphosphate (Bioline, London, UK), 2 mM $MgCl_2$ (Bioline), 0.12 mg ml^{-1} of Bovine serum albumin (Fermentas), 10 pmol of each primer, 1.5 U of *Taq* DNA polymerase (Bioline) and the PCR buffer. The cycling conditions were an initial denaturation at 94°C for 10 min followed by 30 cycles of 94°C for 1 min, 60°C for 1 min, 72°C for 1 min and 30 s and a final 10 min extension at 72°C.

The products from each tagged primer were purified (UltraClean® PCR Clean-Up Kit, Mobio, Carlsbad, CA, USA) and quantified using PicoGreen (Promega, Sunnyvale, CA, USA). Finally, the amplicons of all of the samples were pooled at equimolar concentrations and pyrosequenced using a Roche 454 GS-FLX system (Titanium Chemistry) by GATC (Konstanz, Germany). Raw data and some bioinformatics's treatments described below have been deposited in Dryad (http://datadryad.org/) for public download.

Data processing

The pyrosequencing data, representing 348 422 raw sequence reads, were cleaned by applying PANGEA (Giongo *et al.*, 2010) procedure with a quality threshold (i.e. > 22) and a minimum sequence length of 200 bp. In the following step, reads with forward primer found with at least one mismatch (errors at the beginning of the reads) and with at least one undetermined base (N) were eliminated. The putative chimeras were detected by UCHIME (Edgar *et al.*, 2011) implemented in the USEARCH package (Edgar, 2010). These cleaning procedures and the threshold for delineating an OTU were tested firstly on the *Blastocystis* reads. These sequences in all of the samples were selected and removed of the dataset by the BLAST approach against our reference database (Supporting information S2). Among the reads recovered from this DNA control, 8.2% exactly matched the reference sequence

while 1878 distinct variants were identified, among which 58.5% were singletons. These results corresponded to the rate of errors that occurred during the PCR and pyrosequencing processes. One cluster with USEARCH (Edgar, 2010) was found from the cleaned *Blastocystis* sequences (quality score > 22 and sequence length > 200 bp) originating from a unique strain, if the cut-off for delineating an OTU was defined to 95%. The OTU number and the rare biosphere can be influenced by the threshold for determining an OTU and also by the clustering algorithm used. We compared therefore USEARCH with some algorithms used in the field of the microbial ecology [implemented in MOTHUR (Schloss *et al*., 2009)]: nearest neighbour, furthest neighbour and average neighbour. The results obtained from the *Blastocystis* reads showed that these methods overestimated the richness and the importance of the rare biosphere (results no shown). All the reads were therefore cleaned according to the procedure described and clustered with a cut-off of 95% with USEARCH. After the cleaning step, 304 630 sequences representing 1151 OTUs were selected. By using another denoising method, AmpliconNoise (Quince *et al*., 2011), the number of OTUs was 2.3% higher than the number of OTUs defined by our method.

The OTUs were compared against our reference database (details in Supporting information S2) with USEARCH (Edgar, 2010) and following the taxonomy of its best hit, each sequence is appended to a phyletic group, together with its five best hits. Homologous reads have been then assigned to phyletic groups, they were aligned with the referenced sequences of the corresponding profile using HMMalign (Eddy, 1998) (available in the Dryad data package/alignments). FASTTREE (Price *et al*., 2009) was used to build phylogenetic trees for each phyletic profile with the Jukes-Cantor + Cat model and a bootstrap threshold of 100 (available in the Dryad data package/tree). The package used for this analysis (named PANAM) can be obtained from http://code.google.com/p/panam-phylogenetic-annotation/. After this phylogenetic affiliation, the OTUs belonging to non-picoplanktonic organisms (e.g. metazoan, rhodophyta and streptophyta taxa) were removed. In addition, some biases on the amplification step for one sample have been observed. Indeed, among the 13 038 sequences generated by pyrosequencing on day 38 (24 July), none was affiliated to our internal standard but sequences were rather affiliated to a single perkinsozoan OTU (data not shown). By excluding this sample, a total of 208 319 were therefore selected for studying the small eukaryote dynamics (on average, 9057 reads per sample). For quantitative purposes, the pyrosequencing reads were then normalized on the basis of the *Blastocystis* rDNA added to each sample before the amplification.

According to the definitions of Pedrós-Alió (2006) and Galand and colleagues (2009), OTUs were classified as abundant ($\geq$ 1%) and rare taxa ($\leq$ 0.01%). Between these two classes, a third class for OTUs of intermediate abundance was defined in this study. Furthermore, to estimate the 'original' diversity determined by the pyrosequencing approach, USEARCH and MOTHUR was used to compare the OTUs defined in this study with the sequences of lacustrine small eukaryotes that have been deposited in public databases (sequences for which the specific region amplified in this study, a part of V4–V5, was provided).

The dissimilarity (*D*) index was calculated according to Boucher and colleagues (2006). This is given by $D(t_1,t_2) = 1/2\ \Sigma|xt_1 - xt_2|$, where $\Sigma xt_1 = \Sigma xt_2 = 100$, and xt_1 and xt_2 indicate the relative abundance of a specific OTU at two consecutives dates. The *D*-value ranges from 0 to 100, and it was used to compare samples from one date to another date.

Acknowledgements

We thank J.-C. Hustache and P. Perney for their technical contributions to sampling. This study was supported by financial aids from Cible Région Rhône Alpes programme.

References

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., and Huse, S.M. (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* **4:** e6372.

Amend, A.S., Seifert, K.A., and Bruns, T.D. (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Mol Ecol* **19:** 5555–5565.

Anneville, O., Ginot, V., Druart, J.-C., and Angeli, N. (2002) Long-term study (1974-1998) of seasonal changes in the phytoplankton in Lake Geneva: a multi-table approach. *J Plankton Res* **24:** 993–1008.

Bik, H.M., Sung, W.A.Y., De Ley, P., Baldwin, J.G., Sharma, J., Rocha-Olivares, A., and Thomas, W.K. (2012) Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Mol Ecol* **21:** 1048–1059.

Boucher, D., Jardillier, L., and Debroas, D. (2006) Succession of bacterial community composition over two consecutive years in two aquatic systems: a natural lake and a lake-reservoir. *FEMS Microbiol Ecol* **55:** 79–97.

Brugerolle, G. (2002) *Cryptophagus subtilis*: a new parasite of cryptophytes affiliated with the Perkinsozoa lineage. *Eur J Protistol* **37:** 379–390.

Brugerolle, G. (2003) Apicomplexan parasite *Cryptophagus* renamed *Rastrimonas* gen. nov. *Eur J Protistol* **39:** 101–101.

Campbell, B.J., Yu, L., Heidelberg, J.F., and Kirchman, D.L. (2011) Activity of abundant and rare bacteria in a coastal ocean. *Proc Natl Acad Sci USA* **108:** 12776–12781.

Caron, D.A., and Countway, P.D. (2009) Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquat. Microb Ecol* **57:** 227–238.

Caron, D.A., Countway, P.D., Savai, P., Gast, R.J., Schnetzer, A., Moorthi, S.D., *et al*. (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75:** 5797–5808.

Chambouvet, A., Morin, P., Marie, D., and Guillou, L. (2008) Control of toxic marine dinoflagellate blooms by serial parasitic killers. *Science* **322:** 1254–1257.

Cheung, M.K., Au, C.H., Chu, K.H., Kwan, H.S., and Wong, C.K. (2010) Composition and genetic diversity of picoeu-

karyotes in subtropical coastal waters as revealed by 454 pyrosequencing. *ISME J* **4:** 1053–1059.
Cottrell, M.T., and Suttle, C.A. (1995) Genetic diversity of algal viruses which lyse the photosynthetic picoflagellate *Micromonas pusilla* (Prasinophyceae). *Appl Environ Microbiol* **61:** 3088–3091.
Countway, P.D., Gast, R.J., Savai, P., and Caron, D.A. (2005) Protistan diversity estimates based on 18S rDNA from seawater incubations in the Western North Atlantic. *J Eukaryot Microbiol* **52:** 95–106.
Dawson, S., and Hagen, K. (2009) Mapping the protistan 'rare biosphere'. *J Biol* **8:** 105.
Díez, B., Pedrós-Alió, C., Marsh, T.L., and Massana, R. (2001) Application of denaturing gradient gel electrophoresis (DGGE) to study the diversity of marine picoeukaryotic assemblages and comparison of DGGE with other molecular techniques. *Appl Environ Microbiol* **67:** 2942–2951.
Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* **14:** 755–763.
Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460–2461.
Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194–2200.
Galand, P.E., Casamayor, E.O., Kirchman, D.L., and Lovejoy, C. (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc Natl Acad Sci USA* **106:** 22427–22432.
Gifford, S.M., Sharma, S., Rinta-Kanto, J.M., and Moran, M.A. (2010) Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J* **5:** 461–472.
Giongo, A., Crabb, D.B., Davis-Richardson, A.G., Chauliac, D., Mobberley, J.M., Gano, K.A., *et al*. (2010) PANGEA: pipeline for analysis of next generation amplicons. *ISME J* **4:** 852–861.
Hoshino, T., Yilmaz, L.S., Noguera, D.R., Daims, H., and Wagner, M. (2008) Quantification of target molecules needed to detect microorganisms by fluorescence in situ hybridization (FISH) and catalyzed reporter deposition-FISH. *Appl Environ Microbiol* **74:** 5068–5077.
Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., *et al*. (2007) Microbial population structures in the deep marine biosphere. *Science* **318:** 97–100.
Huse, S.M., Welch, D.M., Morrison, H.G., and Sogin, M.L. (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12:** 1889–1898.
Jacquet, S., Lennon, J.F., Marie, D., and Vaulot, D. (1998) Picoplankton population dynamics in coastal waters of the northwestern Mediterranean Sea. *Limnol Oceanogr* **43:** 1916–1931.
Jobard, M., Rasconi, S., and Sime-Ngando, T. (2010) Diversity and functions of microscopic fungi: a missing component in pelagic food webs. *Aquat Sci* **72:** 255–268.
Jones, M.D.M., Forn, I., Gadelha, C., Egan, M.J., Bass, D., Massana, R., and Richards, T.A. (2011) Discovery of novel intermediate forms redefines the fungal tree of life. *Nature* **474:** 200–203.
Jones, S.E., and Lennon, J.T. (2010) Dormancy contributes to the maintenance of microbial diversity. *Proc Natl Acad Sci USA* **107:** 5881–5886.
Kock, D., and Schippers, A. (2008) Quantitative microbial community analysis of three different sulfidic mine tailing dumps generating acid mine drainage. *Appl Environ Microbiol* **74:** 5211–5219.
Larsson, U., and Hagström, A. (1979) Phytoplankton exudate release as an energy source for the growth of pelagic bacteria. *Mar Biol* **52:** 199–206.
Lefèvre, E., Bardot, C., Noel, C., Carrias, J.-F., Viscogliosi, E., Amblard, C., *et al*. (2007) Unveiling fungal zooflagellates as members of freshwater picoeukaryotes: evidence from a molecular diversity study in a deep meromictic lake. *Environ Microbiol* **9:** 61–71.
Lefèvre, E., Roussel, B., Amblard, C., and Sime-Ngando, T. (2008) The molecular diversity of freshwater picoeukaryotes reveals high occurrence of putative parasitoids in the plankton. *PLoS ONE* **3:** e2324. doi:10.1371/journal.pone.0002324.
Lefèvre, E., Jobard, M., Venisse, J.S., Bec, A., Kagami, M., Amblard, C., *et al*. (2010) Development of a Real-Time PCR assay for quantitative assessment of uncultured freshwater zoosporic fungi. *J Microbiol Methods* **81:** 69–76.
Lefranc, M., Thenot, A., Lepère, C., and Debroas, D. (2005) Genetic diversity of small eukaryotes in lakes differing by their trophic status. *Appl Environ Microbiol* **71:** 5935–5942.
Lepère, C., Domaizon, I., and Debroas, D. (2008) Unexpected importance of potential parasites in the composition of the freshwater small eukaryote community. *Appl Environ Microbiol* **74:** 2940–2949.
Lepère, C., Masquelier, S., Mangot, J.-F., Debroas, D., and Domaizon, I. (2010) Vertical structure of small eukaryotes in three lakes that differ by their trophic status: a quantitative approach. *ISME J* **4:** 1509–1519.
Lindström, E.S., and Langenheder, S. (2011) Local and regional factors influencing bacterial community assembly. *Environ Microbiol Rep* **4:** 1–9.
Liu, H., Probert, I., Uitz, J., Claustre, H., Aris-Brosou, S., Frada, M., *et al*. (2009) Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proc Natl Acad Sci USA* **106:** 12803–12808.
López-García, P., Rodriguez-Valera, F., Pedrós-Alió, C., and Moreira, D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409:** 603–607.
Lozupone, C.A., and Klein, D.A. (2002) Molecular and cultural assessment of chytrid and *Spizellomyces* populations in grassland soils. *Mycologia* **94:** 411–420.
Mangot, J.-F., Lepère, C., Bouvier, C., Debroas, D., and Domaizon, I. (2009) Community structure and dynamics of small eukaryotes targeted by new oligonucleotide probes: new insight into the lacustrine microbial food web. *Appl Environ Microbiol* **75:** 6373–6381.
Mangot, J.-F., Debroas, D., and Domaizon, I. (2011) Perkinsozoa, a well-known marine protozoan flagellate parasite group, newly identified in lacustrine systems: a review. *Hydrobiologia* **659:** 37–48.
Medinger, R., Nolte, V., Pandey, R.V., Jost, S., Ottenwälder, B., Schlötterer, C., and Boenigk, J. (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol Ecol* **19:** 32–40.

Mironova, E., Telesh, I., and Skarlato, S. (2012) Diversity and seasonality in structure of ciliate communities in the Neva Estuary (Baltic Sea). *J Plankton Res* **34:** 208–220.

Monchy, S., Sanciu, G., Jobard, M., Rasconin, S., Gerphagnon, M., Chabé, M., *et al*. (2011) Exploring and quantifying fungal diversity in freshwater lake ecosystems using rDNA cloning/sequencing and SSU tag pyrosequencing. *Environ Microbiol* **13:** 1433–1455.

Nolte, V., Pandey, R.V., Jost, S., Medinger, R., Ottenwälder, B., Boenigk, J., *et al*. (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol Ecol* **19:** 2908–2915.

Not, F., Latasa, M., Scharek, R., Viprey, M., Karleskind, P., Balague, V., *et al*. (2008) Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep-Sea Res, Part I* **55:** 1456–1473.

Not, F., del Campo, J., Balague, V., de Vargas, C., and Massana, R. (2009) New insights into the diversity of marine picoeukaryotes. *PLoS ONE* **4:** e7143. doi:10.1371/journal.pone.0007143.

Park, M.G., Yih, W., and Coats, D.W. (2004) Parasites and phytoplankton, with special emphasis on dinoflagellate infections. *J Eukaryot Microbiol* **51:** 145–155.

Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? *Trends Microbiol* **14:** 257–263.

Pelletier, J.-P., and Orand, A. (1978) Appareil de prélèvement d'un échantillon dans un fluide. Patent number 76.08579.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26:** 1641–1650.

Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., *et al*. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6:** 639–641.

Quince, C., Lanzen, A., Davenport, R., and Turnbaugh, P. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12:** 38.

Rasconi, S., Jobard, M., and Sime-Ngando, T. (2011) Parasitic fungi of phytoplankton: ecological roles and implications for microbial food webs. *Aquat Microb Ecol* **62:** 123–137.

Reckermann, M., and Veldhuis, M.J.W. (1997) Trophic interactions between picophytoplankton and micro- and nanozooplankton in the western Arabian Sea during the NE monsoon 1993. *Aquat Microb Ecol* **12:** 263–273.

Reeder, J., and Knight, R. (2009) The 'rare biosphere': a reality check. *Nat Methods* **6:** 636–637.

Richards, T.A., Vepritskiy, A.A., Gouliamova, D.E., and Nierzwicki-Bauer, S.A. (2005) The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. *Environ Microbiol* **7:** 1413–1425.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al*. (2009) Introducing MOTHUR: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75:** 7537–7541.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., *et al*. (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103:** 12115–12120.

Stoeck, T., Behnke, A., Christen, R., Amaral-Zettler, L., Rodriguez-Mora, M., Chistoserdov, A., *et al*. (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol* **7:** 72.

Tarbe, A.L., Stenuite, S., Balagu, X.E., Sinyinza, D., Descy, J.-P., *et al*. (2011) Molecular characterisation of the small eukaryote community in a tropical Great Lake (Lake Tanganyika, East Africa). *Aquat Microb Ecol* **62:** 177–190.

Urich, T., Lanzen, A., Qi, J., Huson, D.H., Schleper, C., and Schuster, S.C. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* **3:** e2527. doi:10.1371/journal.pone.0002527.

Vaulot, D., and Marie, D. (1999) Diel variability of photosynthetic picoplankton in the equatorial Pacific. *J Geophys Res* **104:** 3297–3310.

Vigil, P., Countway, P.D., Rose, J., Lonsdale, D.J., Gobler, C.J., and Caron, D.A. (2009) Rapid shifts in dominant taxa among microbial eukaryotes in estuarine ecosystems. *Aquat Microb Ecol* **54:** 83–100.

Zhao, B., Chen, M., Sun, Y., Yang, J., and Chen, F. (2011) Genetic diversity of picoeukaryotes in eight lakes differing in trophic status. *Can J Microbiol* **57:** 115–126.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., *et al*. (2011) Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J* **5:** 1303–1313.

Zhu, F., Massana, R., Not, F., Marie, D., and Vaulot, D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52:** 79–92.

Supporting information

Additional Supporting Information may be found in the online version of this article:

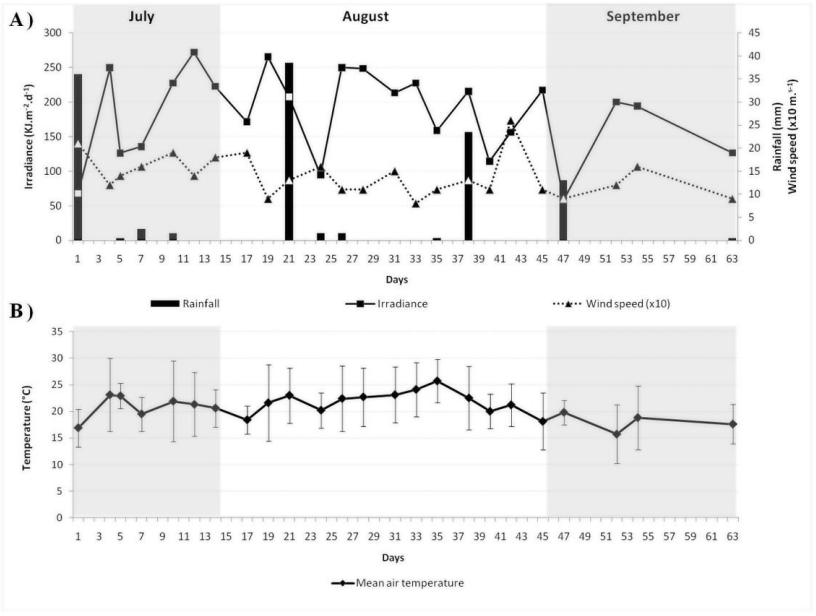
Fig. S1. Main climatologically parameters measured at the permanent meteorological station of the INRA's Thonon station over the 63 days of experimentation (from 17 July to 18 September 2009). Variations of precipitation (mm), wind speed ($\times 10$ m s^{-1}) and solar radiation (KJ m^{-2} day^{-1}) during the study period.

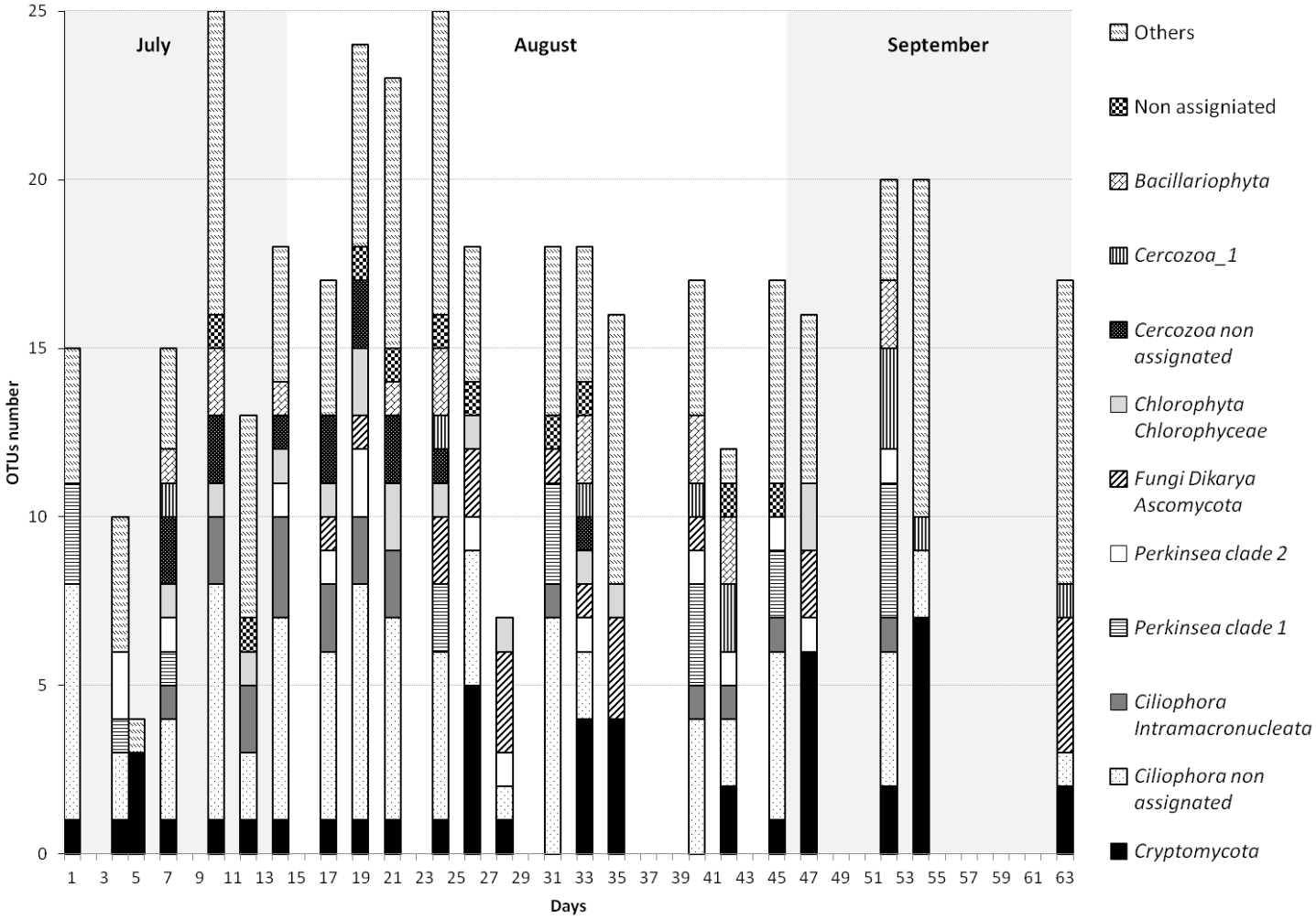
Fig. S2. Variations of the abundant small eukaryotic groups (>1% of total sequences per sample) over the 63 days of experimentation in Lake Geneva (0–20 m).

Table S1. Oligonucleotide probes used in this study.

Supporting information S1. TSA-FISH protocol.

Supporting information S2. Pyrosequencing data analysis.


A)
July
August
September
Irradiance (KJ.m⁻².d⁻¹)
Rainfall (mm)
Wind speed (x10 m.s⁻¹)
Days
Rainfall
Irradiance
Wind speed (x10)
B)
Temperature (°C)
Days
Mean air temperature


July
August
September
OTUs number
Days
0
5
10
15
20
25
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63
Others
Non assigniated
Bacillariophyta
Cercozoa_1
Cercozoa non assignated
Chlorophyta Chlorophyceae
Fungi Dikarya Ascomycota
Perkinsea clade 2
Perkinsea clade 1
Ciliophora Intramacronucleata
Ciliophora non assignated
Cryptomycota

Probe	Sequence (5'-3')	*E..coli* positions	Specificity	Reference
EUK1209R	GGG CAT CAC AGA CCT G	1194-1210	Eukaryota	(Giovannoni *et al.*, 1988)
CERC_02	AAT ACG AGC ACC CCC AAC	663-681	Cercozoa	(Mangot *et al.*, 2009)
NCHLO01	GCT CCA CTC CTG GTG GTG	931-949	Non Chlorophyceae	(Simon *et al.*, 1995)
CHLO02	CTT CGA GCC CCC AAC TTT	765-783	Chlorophyceae	(Simon *et al.*, 2000)
LKM11_01	TAC TGT CAC TAC CTC GCC	384-402	Cryptomycota	(Mangot *et al.*, 2009)
LKM11_02	TGG TCC TCA AAC CAA C	651-667	Cryptomycota	(Mangot *et al.*, 2009)
MY1574	TCC TCG TTG AAG AGC	1324-1339	Fungi (Eumycota)[1]	(Baschien *et al.*, 2008)
PERKIN_01	GAG GAT GCC TCG GTC AA	638-655	Perkinsozoa	(Mangot *et al.*, 2009)
PERKIN_02	GCC AAA CAT TG T ACT GCG	651-669	Perkinsozoa	(Mangot *et al.*, 2009)

[1] Except Cryptomycota.

Supporting information S1

TSA-FISH Protocol

Hybridization conditions for the FISH techniques were applied as described by Amann *et al*. (1995) and Not *et al.* (2002). Briefly, the hybridization filters were covered with a hybridization buffer (40% deionized formamide, 0.9 M NaCl, 20 mM Tris-HCl [pH 7.5], 0.01% sodium dodecyl sulfate, 10% blocking reagent [Roche Diagnostics/Boehringer]) and oligonucleotide probes labeled with horseradish peroxidase (50 ng μl^{-1}) (listed in Table S1). The mixture was left to hybridize at 35°C for 3 h. After two successive 20 min rinses at 37°C in a wash buffer (56 mM NaCl, 5 mM EDTA, 0.01% sodium dodecyl sulfate, 20 mM Tris-HCl [pH 7.5]), samples were equilibrated in TNT buffer (7% Tween 20, 150 mM NaCl, 100 mM Tris-HCl [pH 7.5]) at room temperature for 15 min. Tyramide amplification was performed for 30 min at room temperature in the dark in TSA mix, a mixture (1:1) of 40% dextran sulfate (Sigma-Aldrich) and 1X amplification diluent (Perkin-Elmer LAS), which provide enhanced sensitivity, to which is added fluorescein isothiocyanate coupled with tyramide (1X; Perkin-Elmer LAS) (1:50). Filters were then transferred through two successive 5 ml TNT buffer baths at 55°C for 20 min each to stop the enzymatic reaction and remove the dextran sulfate. Filters were mounted in a mixture of antifading oil AF1 (Citifluor, Biovalley, Conches, France) containing 10 $\mu g\ ml^{-1}$ of propidium iodide (Sigma-Aldrich).

Supporting information S2

Pyrosequencing data analysis

Reference database

Experimental sequences analyzed through pyrosequencing were compared with a dedicated database of reference sequences extracted from the SSURef 108 database from the SILVA database project, which offers taxonomic information, quality assessment and a curated alignment of SSU rRNA sequences (Pruesse *et al*., 2007). The database only includes sequences with more than 1 200 bp, quality score > 75%, and pintail value > 50 according to the SILVA classification. In this implementation, sequences that belong to the *Eukarya* domain were kept, corresponding to 15 419 sequences from microeukaryotes. Sequences from pluricellular organisms (*Metazoa*, *Rhodophyta* and *Streptophyta*), plus one bacterial sequence and one archaeal sequence were added to allow the detection of experimental sequences affiliated to them.

To speed up the phylogenetic processing, the reference alignments were split into 24 phyletic groups. For each phyletic group, an outgroup containing one sequence from each of the other phyletic groups plus two metazoan sequences were added to the alignment to root the phyletic tree to be produced, and to specify the relatedness of early diverging sequences from the root of the group. The sequences from each phyletic group together with the outgroup sequences were retrieved from the SILVA alignment using ARB, and then trimmed to remove vertical gaps. Besides these files, an HMM profile was built from each of them using HMMbuild from the HMMER package (Eddy, 1998). A taxonomy file containing the taxonomy of each sequence of the reference database was also generated.

Processed sequences

1- Cleaning procedures:

By using PANGEA functionalities (Giongo *et al*., 2010), short sequences (< 200 bp) and sequences with low quality (≤ 22) were removed. After applying quality score (> 22), read length, tags and adaptators trimming, 315 367 were kept (*i.d.* 9.49% removed). By eliminating the reads with forward primer found with at least one mismatch (errors at the beginning of the reads), 309 878 were conserved (*i.d*. 11.06% removed). By removing the reads with at least one undetermined base, 304 792 were preserved (*i.d*. 12.52% removed). Finally, we used UCHIME (Edgar *et al*., 2011) for detecting chimera corresponding to 162 reads; 304 630 sequences were kept at this step.

2- Cleaned reads were then clustered with UCLUST included in USEARCH (Edgar, 2010).

3- The OTUs were compared against the reference database with USEARCH .Then, following the taxonomy of its best hit, each sequence was appended to a phyletic group, together with its five best hits. The query sequences were sorted according to their assignment.

4- Homologous reads have been then assigned to phyletic groups, they were aligned with the referenced sequences of the corresponding profile using HMMalign . Next, a phylogenetic tree was built for each phyletic profile, using FASTTREE (Price *et al*., 2009), with the Jukes-Cantor + Cat model and a bootstrap threshold of 100.

5- The trees were parsed to generate files containing the taxonomy of the inserted sequences. Taxonomy assessment was inferred by nearest neighbor. For each query sequence, all the nodes containing it were scanned from the most recent to the deepest. The closest neighbor was defined as the first referenced sequence starting from the lowest node. The query

sequence acquires the complete taxonomy of its closest neighbor.

6- In the last step, the pipeline describes the monophyletic clusters with all information enabling experts in the field to define a putative environmental clade: a bootstrap value, a list of all the experimental sequences affiliated to it and the nearest reference neighbor together with its taxonomy. This analysis allowed us to verify the affiliation of OTUs to the lacustrine environmental clades.

Accuracy of the phylogenetic affiliation processed by PANAM

The reliability of PANAM was evaluated using an amplicon simulation to generate pseudo-reads by clipping the 5 × 1 000 full-length sequences from PANAM database. This affiliation was compared to BLAST used to affiliate pyrosequencing reads from eukaryotes (*e.g.* Stoeck *et al*., 2009, Nolte *et al*., 2010). Our results show that the taxonomic affiliation given by PANAM is always better than BLAST (Fig. 1).

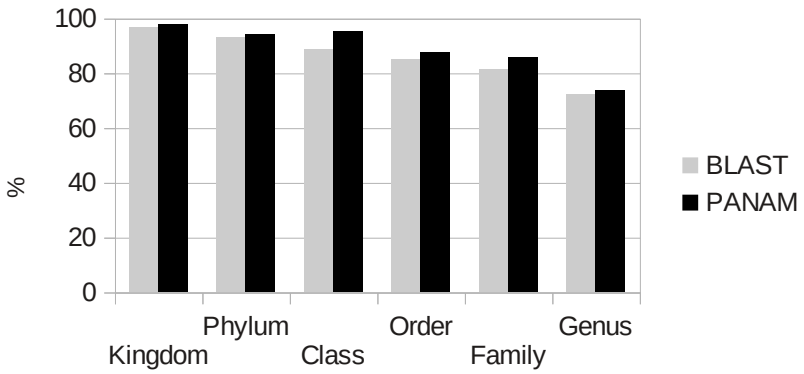


Fig. 1: Accuracy of the phylogenetic affiliation by PANAM compared to the BLAST. 1 000 sequences were randomly picked from the reference database and removed from it for the simulations. Simulations were repeated five times and the standard variation is less than 0.03.

The package used for this analysis (named PANAM) can be obtained from http://code.google.com/p/panam-phylogenetic-annotation/. It comprises the reference sequences database, the taxonomy file and reference profile alignments.

References

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and *in-situ* detection of individual microbial-cells without cultivation. *Microbiol Rev* **59:** 143-169.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* **14**: 755-763.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460-2461.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194-2200.

Giongo, A., Crabb, D.B., Davis-Richardson, A.G., Chauliac, D., Mobberley, J.M., Gano, K.A., Mukherjee, N., Casella, G., Roesch, L.F., Walts, B. *et al*. (2010) PANGEA: pipeline for analysis of next generation amplicons. *ISME J* **4:** 852-861.

Nolte, V., Pandey, R.V., Jost, S., Medinger, R., Ottenwälder, B., Boenigk, J. *et al.* (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol Ecol* **19:** 2908-2915.

Not F, Simon N, Biegala IC, and Vaulot D. (2002). Application of fluorescent *in situ* hybridization coupled with tyramide signal amplification (FISH-TSA) to assess eukaryotic picoplankton composition. *Aquat Microb Ecol* **28:** 157-166.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641-1650.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., and Peplies, J. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188-7196.

Stoeck, T., Behnke, A., Christen, R., Amaral-Zettler, L., Rodriguez-Mora, M., Chistoserdov, A. *et al.* (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol* **7:** 72.

Chapitre 4

Discussion

L'utilisation du séquençage à haut débit en écologie microbienne a apporté plusieurs opportunités de recherche, mais également de nouveaux défis, principalement pour ce qui est de l'analyse des données. Si le séquençage massif a permis l'accès à des séquences de micro-organismes jusque-là inconnus et correspondant principalement à la biosphère rare, le traitement de ces données nouvellement acquises pose toujours un défi d'analyse. Dans le contexte de la métagénétique, un défi majeur réside dans l'affiliation taxonomique de gros volumes de données pour déterminer la composition d'une communauté, mais également dans le traitement des biais inhérents à l'utilisation de ces nouvelles techniques. Actuellement, il est bien établi qu'à partir d'une taille d'amplicons de 400 pb, l'affiliation taxonomique par reconstruction phylogénétique est plus robuste que les approches basées sur la similitude et les k-mers (Price et al., 2010). Toutefois, en métagénétique l'assignation taxonomique de centaine de milliers de séquences relativement courtes générées par le séquençage massif est principalement réalisée par le classificateur RDP (Wang et al., 2007) ; BLAST (Altschul et al., 1997) ou UCLUST, implémentés dans QIIME (Caporaso et al., 2010) ou Mothur (Schloss et al., 2009), alors que les phylogénies sont rarement utilisées.

Dans ce travail de thèse, nous avons implémenté une approche phylogénétique (PANAM) pour l'assignation taxonomique et la description phylogénétique de gros volumes de données, qui sera mis à la disposition de la communauté scientifique à travers un service web dédié à l'analyse des données de métagénétique pour la description de la diversité microbienne. Cette approche a été utilisée pour étudier la composition taxonomique de deux modèles, les *Archaea* en milieu marin et les picoeucaryotes en milieux lacustres. L'application de PANAM dans le cadre de nos trois études (chapitres 2 et 3) a, d'une part confirmé la robustesse de l'assignation taxonomique de cette approche par rapport aux autres à partir d'une taille d'amplicon contenant un signal phylogénétique assez résolutif, et d'autre part elle a montré la capacité de retracer les clades environnementaux et à en identifier de nouveaux. De plus, l'utilisation d'un standard interne dans l'une de nos études (Chapitre 3, article 4) nous a permis de compléter les stratégies de nettoyage et de clusterisation employées dans PANAM pour une meilleure approximation de la richesse.

4.1 PANAM et l'affiliation taxonomique

4.1.1 Impact de la région et de la taille de l'amplicon sur les assignations

La comparaison des différentes approches d'annotation dans l'étude présentée dans le chapitre 2 a permis de montrer un net avantage pour les approches phylogénétiques sur les séquences complètes et les fragments à partir de 400 pb. En effet, alors que l'utilisation des approches phylogénétiques en métagénétique était initialement limitée à cause du faible signal porté par les courts fragments de l'ADNr générés par les techniques de séquençage (≈150 pb), les progrès récents dans ces techniques permettent de générer des fragments de tailles convenables pour les analyses phylogénétiques (Jeraldo et al., 2011). Dans notre étude, nous avons comparé l'impact de la taille des fragments de l'ADNr 18S sur la fiabilité des différentes approches. S'il est clair que toutes les approches d'affiliation gagnent en précision en augmentant la taille des amplicons, ce gain est différent en fonction de l'approche. En effet, alors qu'à 200 pb BLAST a la spécificité [1] la plus élevée pour toutes les régions testées, les approches phylogénétiques sont les plus performantes à 400 pb, indiquant un accroissement plus important de la spécificité entre 200 pb et 400 pb pour ces approches. Ces résultats sont en accord avec les résultats de Liu et al. (2008), qui ont comparé différentes approches d'assignation taxonomique sur des fragments de l'ADNr 16S de 100, 250 et 400 pb, et ont conclu que les méthodes basées sur les phylogénies sont les plus sensibles à la taille des fragments. Outre la taille des fragments, les méthodes phylogénétiques sont plus sensibles à la région amplifiée. L'étude de Liu et al. (2008) montre que l'augmentation de la taille des régions avec les résolutions phylogénétiques les plus faibles n'améliore pas leur précision par rapport aux autres régions. Ces résultats sont similaires à nos conclusions pour les séquences de l'ADNr 18S (Chapitre 2, article 1). Selon Schloss, (2010), si l'impact de la région amplifiée de l'ADNr 16S sur les estimations de la diversité dépend de l'approche utilisée, c'est essentiellement dû aux différents taux d'évolution entre les régions variables ; et à la différence entre la diver-

1. le pourcentage de groupes taxonomiques correctement assignés par rapport aux groupes taxonomiques détectés.

sité génétique des séquences, utilisée pour les assignations par BLAST par exemple, et la diversité phylogénétique calculée en additionnant les branches de l'arbre. En métagénétique, le choix de la région à amplifier est régi par des contraintes d'ordre technique et des contraintes liées à l'analyse. Techniquement, les régions amplifiées doivent être délimitées par des zones hautement conservées permettant la fixation d'amorces dites « universelles » ; elles doivent avoir des tailles adaptées aux techniques de séquençage disponibles et elles doivent pouvoir générer des séquences de bonne qualité (e.g., absence des homopolymères pour le pyroséquençage). D'un point de vue de l'analyse, la région amplifiée doit être fortement représentée dans les bases de référence (e.g., la région v9 est faiblement présente dans les bases de données publiques) ; elle doit pouvoir restituer l'information portée par la séquence complète ; et elle doit couvrir la diversité génétique. Dans notre étude sur les *Archaea* (Chapitre 3, article 3), notre choix de la région v3-v5 a été guidé par des études antérieures sur l'ADNr 16S, notamment l'étude de Kim et al. (2011). Les amplicons générés nous ont permis de retracer les clades MGI.A et MGI.B, définis dans la littérature à partir de séquences complètes, indiquant ainsi que l'information portée par la région v3-v5 sur les séquences de l'ADNr 16S des *Archaea* restitue l'information phylogénétique des séquences complètes. En ce qui concerne les séquences de l'ADNr 18S, nos simulations ont montré qu'à 400 pb, la région v8-v9 affiche la meilleure valeur en terme de spécificité, toutefois, étant donnée la faible représentativité de cette région dans la base utilisée, et pour l'étude réalisée sur les picoeucaryotes lacustres notre choix s'est porté sur la région v4 bien qu'elle ait une spécificité plus faible (76.8% contre 79.2%). Afin de palier à la vision partielle de la diversité qui peut être produite par certaines régions, une des solutions serait l'utilisation de plusieurs couples d'amorces, tel que proposé par Stoeck et al. (2010) dans leur étude sur la diversité des communautés des protistes dans des eaux marines. Dans cette étude, les régions variables v4 et v9 ont été utilisées et, alors que la région v9 délimitait les groupes à des niveaux taxonomiques élevés, la région v4 permettait de différentier les espèces proches. Selon les auteurs, ces deux régions auraient des avantages différents en fonction des groupes taxonomiques ciblés et de la question biologique du départ. Ces résultats sont en accord avec notre étude sur les différentes régions hypervariables de l'ADNr 18S (chapitre 2, article 1), où il apparaît que selon le groupe taxonomique, les régions v4 et v9 ne possèdent pas la même spécificité au niveau du genre.

4.1.2 Identification des clades

Le pipeline QIIME, largement utilisé par la communauté des écologistes microbiens, génère les phylogénies soit *de novo*, avec uniquement les séquences représentatives des OTUs en vue d'une estimation de la diversité béta via l'utilisation d'UNIFRAC (Lozupone and Knight, 2005, Lozupone et al., 2006), ou alors il se base sur une phylogénie construite *a priori* sur un ensemble de séquences de référence, et regroupe en OTUs les séquences de cette phylogénie avec les séquences expérimentales afin d'évaluer leurs relations phylogénétiques. La première stratégie permet d'identifier des clades uniquement parmi les séquences étudiées sans évaluer leur proximité avec les espèces connues ; alors que la deuxième ne permet d'identifier que les séquences ayant des homologues proches dans la phylogénie de référence utilisée. L'outil que nous avons développé (PANAM), permet de comparer les séquences requêtes avec une base de référence, de sélectionner un sous ensemble des séquences de référence homologues aux séquences requêtes, et de réévaluer leurs relations en générant une phylogénie *de novo*. Ainsi, l'utilisation de PANAM sur les données issues du pyroséquençage nous a permis de mettre en évidence parmi les *Chlorophyceae*, un clade des Mamiellales qui n'a encore jamais été décrit dans les lacs par les méthodes de séquençage et d'annotation classiques, et dont la détection requiert, selon Marin and Melkonian (2010), l'utilisation d'approches moléculaires différentes. De la même manière, l'application de PANAM sur les communautés des *Archaea* (Chapitre 3, article 3) nous a permis d'affilier des OTUs à des clades décrits dans la littérature (e.g. MGI.A et MGI.B (Galand et al., 2010)), mais également à deux nouveaux clades (C et D) dans la phylogénie des MGI et qui sont constitués exclusivement d'OTUs rares. Une affiliation de ces séquences par une approche de similitude ou probabiliste aurait reflété une appartenance au même groupe taxonomique des MGI sans décrire leur diversité phylogénétique. De plus, la comparaison des OTUs du clade C avec les bases publiques par BLAST les affilie toutes à des séquences environnementales partielles, et la première séquence de l'organisme connu le plus proche (U51469, *Cenarchaeum symbiosum*) ressort à des scores de similitude entre 92% et 95%. En revanche, l'analyse de la structure phylogénétique de ces quatre clades à la lumière des données environnementales, a montré des profils d'activité qui diffèrent selon le clade et en fonction de la saison. Cette spécialisation écologique suggère une adaptation des individus du groupe MGI à des conditions environnementales différentes, qui pourrait signifier la présence d'écotypes.

4.1.3 Impact des bases de référence sur les assignations

Selon une étude de Werner et al. (2011) les assignations taxonomiques dépendent du nombre des séquences dans la base de référence ainsi que de la représentativité des espèces. Les affiliations des approches probabilistes et d'alignement sont par conséquent certainement biaisées par l'enrichissement des bases de données publiques par des séquences provenant de milieux intensément décrits dans la littérature par rapport aux séquences des milieux sous échantillonnés. En effet, quand la séquence requête ne possède pas d'homologues proches dans les bases de référence, BLAST produit toujours un alignement et peut inférer une taxonomie erronée, alors que le classificateur RDP lui attribue des affiliations taxonomiques avec de faibles scores de probabilité. Les méthodes phylogénétiques, quant à elles, replacent les séquences requêtes dans un contexte évolutif et leur attribuent une taxonomie dont la précision dépend de la position par rapport aux séquences voisines ; elles s'affranchissent ainsi partiellement des limitations liées à la richesse de la base de référence pour l'identification de nouveaux taxa en les replaçant dans un contexte évolutif. Quand les séquences étudiées possèdent des voisins dans la base de référence, l'assignation taxonomique est faite selon les liens de parenté et leurs positions dans la phylogénie ; et quand un groupe de séquences provient d'un milieu sous échantillonné, il s'insère plus profondément dans l'arbre, indiquant son degré de proximité par rapport aux séquences de référence. Les micro-organismes eucaryotes sont les plus impactés par ce biais à cause de la faible représentativité dans les bases de données des séquences de l'ADNr 18S. A l'heure actuelle, sur la totalité des séquences de la petite sous unité de l'ADNr de la base SILVA (SSURefNR 114), les séquences appartenant aux eucaryotes constituent seulement 11.5% des données. L'ajout de nouvelles séquences, spécifiques aux milieux étudiées améliore la précision de l'annotation taxonomique. En effet, dans notre étude sur les picoeucaryotes (Chapitre 3, article 4), nous avons ré-annoté la base de référence utilisée en ajoutant des séquences de clades spécifiques aux milieux lacustres, et nous avons pu ainsi détecter la présence d'OTUs affiliées aux clades *Perkinsea* clades 1 et 2 et *Cryptophyta* clades 3 et 4, susceptibles de constituer des écotypes différents. Une telle approche pourrait également être utilisée avec BLAST et RDP Classifier.

4.1.4 Le problème de la notion d'espèce en microbiologie

Des séquences hautement similaires de la petite sous unité de l'ADNr peuvent provenir d'organismes génétiquement distincts mais apparentés, d'organismes de la même espèce, ou encore d'un même génome contenant plusieurs copies d'ADNr 16S ou 18S. En écologie microbienne, la définition des OTUs se basant sur les distances génétiques, est souvent controversée du fait de cette variabilité inter et intra spécifique, et de l'absence d'un seuil de similitude universel. Une des méthodes qui a été proposée pour l'approximation de la notion d'espèce en microbiologie est d'intégrer les informations phylogénétiques pour la définition des OTUs (Martin, 2002, Lozupone and Knight, 2005, Powell et al., 2011). Cette solution s'appuie sur le fait que l'absence d'une cohérence phylogénétique dans les OTUs peut résulter en une hétérogénéité écologique (Koeppel and Wu, 2013).

L'ajout d'un standard interne dans notre jeu de données des picoeucaryotes (Chapitre 3, article 4) nous a permis de comparer les OTUs générées par similitude et les groupes monophylétiques. L'affiliation phylogénétique des 1948 séquences nettoyées de *Blastocystis hominis* a résulté en un groupe monophylétique affilié à la même séquence. Un groupe majoritaire avec 99.7% des séquences, et un deuxième groupe avec cinq séquences dont le singleton généré après clusterisation à 95%. Ces résultats sont différents de ceux de Koeppel and Wu (2013) qui tend à démontrer que des OTUs construites à partir des séquences complètes d'ADNr 16S peuvent être polyphylétiques ou paraphylétiques. Dans notre étude, les séquences qui s'affilient avec le singleton, n'affichent pas des profils de mutations particuliers, par contre elles se caractérisent par des tailles réduites (200 à 250 pb), ce qui peut se traduire par un signal phylogénétique moins résolutif que les séquences du clade majoritaire et expliquer leur positionnement dans l'arbre. De plus, alors que l'étude de Koeppel and Wu (2013), préconise l'utilisation de seuils de similitude plus faibles ou des marqueurs moins conservatifs que la petite sous unité de l'ADNr, nous avons restreint dans notre étude les séquences à une région hypervariable (v4-v5), ce qui pourrait expliquer nos résultats différents d'autant plus que les seuils utilisés dans leur étude (97% et 99%) sont plus conservatifs que le seuil appliqué aux séquences de *B. hominis*. Du fait de l'absence d'un seuil de similitude universel, Koeppel and Wu (2013) suggèrent d'utiliser la notion d'écotype comme unité de la mesure de la diversité. Cette notion d'écotype a été par ailleurs évaluée dans notre analyse de la diversité des *Archaea* (Chapitre 3, article 3). Afin d'évaluer les relations évolutives entre les OTUs définies avec

un seuil de 97%, une phylogénie du groupe MGI a été générée. Les clades C et D mis en évidence par rapport à leurs positions phylogénétiques ont montré une activité en hiver en réponse à certains paramètres (e.g., Oxygène, nitrate, Chlorophylle *a*). Enfin, si la notion d'espèce en microbiologie reste discutée, l'application de plusieurs critères intégrant des données d'ordre phylogénétique, génétique et écologique pourrait réduire les biais propres à chaque méthode et fournir une définition qui consolide ces approches pour une meilleure approximation de la richesse.

4.2 PANAM et les biais de l'estimation de la richesse et de la diversité

De nombreuses études ont été réalisées pour évaluer les biais découlant de l'utilisation de la métagénétique dans la description de la richesse et de la diversité microbiennes (Huse et al., 2007, Quince et al., 2009, Kunin and Hugenholtz, 2010). En effet, la description de la diversité peut être faussée par des estimations erronées de l'abondance relative des espèces à cause d'un effort de séquençage différentiel ou de l'hétérogénéité du nombre des copies du gène ribosomique entre taxa. D'autre part, des biais dans l'estimation de la richesse d'une communauté microbienne peuvent être la conséquence d'un polymorphisme intra spécifique de ces mêmes opérons, du choix du seuil de clusterisation ou d'un traitement inadapté des erreurs dans la génération des séquences (PCR et séquençage). Si certains de ces biais sont d'ordre méthodologique et peuvent être réduits par un traitement bio-informatique adapté, d'autres sont d'ordre biologique et leur résolution dépasse le traitement bio-informatique et nécessite des expérimentations supplémentaires.

4.2.1 Les biais dans l'estimation des indices de richesse et de diversité

En métagénétique, une copie de l'ADNr 16S ou 18S représente un individu, et la capacité à prédire la prévalence d'un micro-organisme dans un environnement repose sur l'abondance des séquences qui lui sont affiliées dans l'échantillon. Or, cette abondance relative dépend d'une part de l'effort de séquençage, et d'autre part du nombre de copies du gène ribosomique dans un génome. En effet, la profondeur de séquençage étant hétéro-

gène parmi les échantillons, les tailles des librairies résultantes ne sont pas égales, ce qui biaise la comparaison des indices qui se basent sur l'abondance relative et les effectifs des échantillons (Gihring et al., 2012). Une des solutions pour comparer la richesse et la diversité de différents échantillons est d'égaliser les effectifs des librairies par normalisation. Cette normalisation peut être réalisée en multipliant les effectifs des échantillons par un facteur de correction afin de les ramener à la même valeur avant de calculer les indices (e.g, Bowen et al. (2012)). La deuxième approche, consiste à tirer aléatoirement le même nombre de séquences dans chaque échantillon avant clusterisation. Cette dernière stratégie a été implémentée dans PANAM à deux niveaux. La normalisation *a priori*, réalisée après nettoyage des séquences brutes en vue de comparer les indices de richesse et de diversité ainsi que les courbes de raréfaction des différents échantillons. L'application de cette normalisation dans l'étude présentée dans l'article 1 (chapitre 2) nous a permis de réévaluer la richesse des lacs. Alors qu'avant la normalisation, les effectifs des séquences variaient entre 3759 (lac Bourget) et 17092 séquences (lac Anterne), avec l'indice Chao1 qui désignait le lac Villerest comme étant le plus riche (Chao1=482.3), après la normalisation par rapport à la taille de l'échantillon le plus faible (3759), le lac Bourget (Chao1=436.2) affichait plus de richesse que le lac Villerest (Chao1=399.5). Un des problèmes de cette stratégie est qu'elle ne prend pas en compte l'affiliation taxonomique et elle peut être biaisée par des amorces dont la spécificité ne serait que partielle. En effet, dans l'étude sur les *Archaea* (article 3, chapitre 3), alors que les effectifs des échantillons variaient de 973 à 20441 après le nettoyage des séquences, ces effectifs variaient de 12 à 7910 lectures après affiliation taxonomique étant donné que 70% des séquences a été écarté car assigné à des bactéries. Ainsi, la deuxième normalisation implémentée dans PANAM est réalisée après l'affiliation taxonomique des séquences, et permet de calculer les indices de diversité et de richesse par groupe taxonomique sur des échantillons avec le même nombre de séquences affiliées aux groupes d'intérêt. Le premier ré-échantillonnage, ne tenant pas compte de la taxonomie des séquences, tend à comparer de manière générale la richesse et la diversité des échantillons ; alors que le deuxième permet de normaliser uniquement par rapport aux séquences affiliées aux organismes d'intérêt.

Le deuxième biais affectant l'estimation de l'abondance relative, est celui associé à la variation du nombre de copies de l'ADNr 16S/18S dans les génomes. En effet, cette varia-

bilité inter spécifique du nombre de copies peut aller d'une à 15 copies pour les bactéries (Acinas et al., 2004), et jusqu'à des centaines de milliers chez certains eucaryotes (Gong et al., 2013). Dans une étude de Medinger et al. (2010), les auteurs ont comparé les estimations de l'abondance relative inférées à partir de la morphologie et des données NGS des alvéolées (*Ciliophora* et *Dinophyceae*), et ont montré une surestimation de l'abondance dans les librairies de séquences. Afin de palier à ce biais, ces auteurs préconisent l'utilisation d'amorces spécifiques pour cibler un groupe taxonomique, ou de restreindre les analyses à des groupes taxonomiques avec le même nombre de répétitions. Ainsi, dans PANAM, les indices de richesse et de diversité sont calculés par rang taxonomique (Phylum et Classe), en supposant que le nombre de copies du gène ribosomique est plus homogène au sein d'un groupe taxonomique.

4.2.2 Les biais dans l'estimation de la richesse

En métagénétique, il est commun de regrouper les séquences dans des OTUs en se basant sur les distances entre les séquences. De nombreuses études ont montré que les erreurs de séquençage résultent en une surestimation de la richesse à cause de la génération à de faibles abondances de faux positifs (Quince et al., 2009). De plus, outre les méthodes de clusterisation (e.g., algorithme, distance), les paramètres à appliquer pour une meilleure approximation de la notion d'espèce dépendent également des conditions de l'expérimentation (e.g., organismes étudiés ; région amplifiée ; qualité de l'ADN utilisé pour les amplifications ; techniques de séquençage ..), d'où la nécessité de la mise en place d'une stratégie de nettoyage et de clusterisation adaptée à ces conditions. Une des méthodes utilisée pour quantifier la surestimation de la richesse, est la confrontation de la diversité génétique et de la diversité phénotypique. C'est par exemple la méthode employée par Bachy et al. (2012) afin de comparer la diversité retrouvée après séquençage et estimer les biais d'amplification et de séquençage. Toutefois, l'utilisation de cette approche limite son application aux espèces microbiennes avec des caractéristiques observables permettant de les différentier facilement des autres espèces. Or, parmi les microorganismes, seuls les protistes possèdent des traits morphologiques distincts. Cependant, la variation génétique au sein des gènes codant la petite sous unité de l'ARNr reflète l'accumulation de mutations neutres et n'est pas forcément corrélée avec une différentiation phénotypique (Fenchel, 2005). De plus, cette approche n'inclue pas les biais inhérents à l'identification de

cette population « contrôle » (e.g., spécificité des amorces), et nécessite une connaissance *a priori* de sa diversité génétique. Enfin, elle suppose que la notion d'espèce morphologique prévaudrait sur la notion d'espèce génétique.

Dans l'étude sur les picoeucaryotes (Chapitre 3, article 4), nous avons procédé à l'ajout d'un étalon interne aux échantillons avant amplification et séquençage. Cette méthode a été appliquée en métatranscriptomique par Gifford et al. (2010) afin d'inférer le nombre de transcrits absolu à partir du nombre de transcrits séquencés. Une autre étude menée en même temps que notre expérimentation, celle de Zhou et al. (2011), avait introduit de l'ADN contrôle dans les échantillons avant l'amplification par PCR afin d'évaluer la reproductibilité des expériences impliquant le séquençage d'amplicons. Ainsi, ce standard interne a été utilisé dans le cadre de cette étude pour : i) normaliser les effectifs des échantillons en comparant ses proportions avant et après amplification, et avoir une corrélation entre ces effectifs corrigés et le nombre de copies d'ADNr 18S estimé par PCR quantitative ; ii) établir un profil d'erreur propre à cette expérimentation en comparant la variabilité moléculaire à partir des séquences obtenues du standard interne avec sa vraie séquence (un seul clone) ; ii) et calibrer nos choix méthodologiques de nettoyage et de clusterisation en fonction de la part et de la nature de la variabilité introduite dans ce jeu de données par les étapes de PCR et de séquençage. La comparaison des séquences du standard interne avec la séquence originale dans notre jeu de données a montré plusieurs types d'erreurs, avec une prévalence des erreurs de type substitution. Ces erreurs peuvent parvenir à différentes étapes du séquençage, depuis de la génération des amplicons par PCR (principalement des substitutions selon V Wintzingerode et al. (1997)), jusqu'au séquençage et notamment lors de l'interprétation des signaux lumineux (principalement des insertions et des délétions selon Gilles et al. (2011)). L'application du filtre qualité avec un score minimal 23 a permis de diminuer la proportion des séquences avec des substitutions de 70.2% à 65%. Cette faible amélioration de la qualité générale des séquences est due au fait que les substitutions, possèdent des scores de qualité similaires aux scores des bases non mutées (≈35 dans notre étude). En effet, ce type d'erreur est principalement généré pendant la PCR (V Wintzingerode et al., 1997), et dépend donc de la préparation des échantillons, de la qualité de l'ADN et du nombre de cycles d'amplification par PCR. Selon une étude de Gilles et al. (2011), la distribution des types d'erreur entre des plates-formes

utilisant la même technique de séquençage est hétérogène, ce qui complique davantage la mise au point de stratégies, au niveau des séquenceurs, pour réduire ces erreurs. Ainsi, le même standard interne a été séquencé par une deuxième plate-forme de séquençage afin de comparer les profils des erreurs propres à chacune. La figure 4.1 montre les différences des erreurs entre ces deux jeux de séquences. En comparant les séquences brutes, il apparaît que la deuxième plate-forme génère plus d'insertions au niveau des séquences de *Blastocystis hominis* (3.84 insertions par séquence contre 0.58 pour le premier jeu de séquences de la plate-forme 1) ; et moins de substitutions. Bien que le filtre qualité appliqué sur les séquences de standard interne nous ait permis de réduire les erreurs, ces nouveaux résultats montrent qu'il faut adapter le filtre à appliquer selon chaque expérience, dès lors que le taux et la nature des erreurs dépendent de la qualité des amplicons et du séquenceur utilisé. L'étude de Gilles et al. (2011) propose l'utilisation de plusieurs couples d'amorces, qui, associés à la profondeur de séquençage massif, permettrait de compenser partiellement les erreurs et de corriger les erreurs aléatoires sur les séquences.

En ce qui concerne le seuil de clusterisation des séquences eucaryotes (Chapitre 3, article 4), nous avons testé le seuil 95%, défini selon l'étude de Caron and Countway (2009) sur les protistes, et par des simulations *in silico* en fonction de la région amplifiée. A partir des séquences nettoyées, ce seuil génère deux OTUs dont un singleton, alors qu'à partir des séquences brutes il génère 23 OTUs.

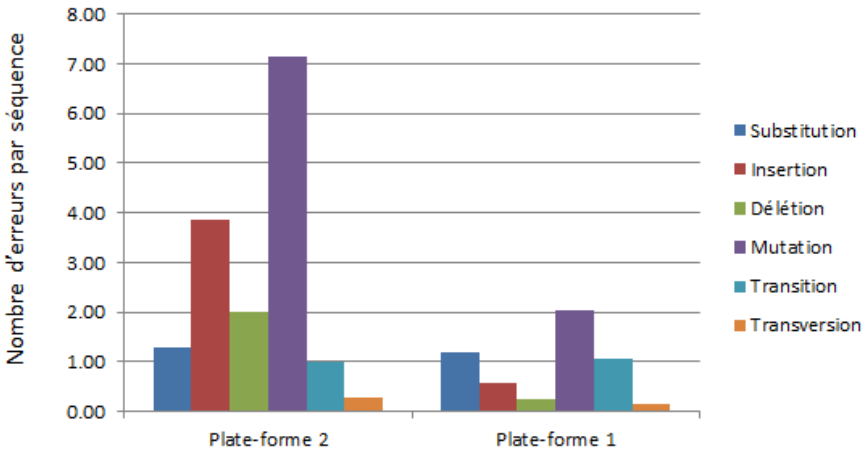


Figure 4.1. Le nombre et la nature des mutations par séquence sur deux jeux de données non nettoyés générés à partir de la séquence de *B. hominis* par deux plates-formes différentes.

Ainsi on peut déduire que, le regroupement des séquences pour la définition d'une unité de la diversité résulte de la combinaison entre le filtre qualité et le seuil de clusterisation. Par ailleurs, certaines études préconisent l'utilisation des seuils conservatifs (e.g., Zaura et al. (2009)) pour réduire une surestimation de la richesse. Toutefois, il se peut que malgré le filtre de qualité et le seuil de clusterisation, on ne puisse exclure l'existence de séquences ayant accumulé un grand nombre de substitutions et formant de nouveaux clusters, c'est d'ailleurs le cas du singleton résiduel de *Blastocystis hominis* qui a une distance maximale de 30% et une distance minimale de 26% par rapport à l'OTU majoritaire, alors que la distance maximale au sein de cette OTU est de 6%. L'utilisation d'un seuil conservatif pour n'obtenir qu'une seule OTU implique une valeur maximale de 74% pour la clusterisation, seuil qui correspond à un regroupement au rang Phylum. Si certaines séquences artéfactuelles, souvent présentes avec une faible abondance, persistent malgré le nettoyage et la clusterisation, ignorer systématiquement les singletons dans l'analyse de la diversité n'est pas toujours justifié, étant donné que dans certaines études ces OTUs ont été correctement affiliées à des organismes connus, et répondaient aux changements des données environnementales. En effet, la caractérisation des espèces rares en relation avec leur structure phylogénétique et en fonction des conditions environnementales permettrait de différentier les séquences des espèces rares impliquées dans le fonctionnement de l'écosystème, des séquences artéfactuelles qui ne présentent aucune structuration que ce soit en termes de phylogénie ou de réponse aux variables de l'environnement. Enfin, si le principal signal écologique reflété par les patrons de diversité à l'échelle de la communauté n'est pas affecté par l'omission des taxa rares (Agogué et al., 2011, Gobet et al., 2010), leur identification n'en reste pas moins importante pour comprendre leur rôle dans le maintien de la diversité face à des changements globaux.

4.3 Conclusion et Perspectives

Ce travail s'inscrit dans le cadre de la problématique de l'étude de la diversité microbienne dans l'ère du séquençage massif. En effet, l'avènement de ces technologies a entraîné le changement d'échelle des données produites, et par conséquence, les besoins en développements bioinformatiques spécifiques ainsi qu'en ressources informatiques dédiées pour le traitement de ces données sont devenus un point critique de l'exploration

de la biodiversité. Des outils dédiés à l'analyse de la diversité ont été développés et sont rapidement devenus populaires au sein de la communauté des écologistes microbiens (e.g., Mothur (Schloss et al., 2009), QIIME (Caporaso et al., 2010), PyroTagger (Kunin and Hugenholtz, 2010)). Cependant, l'affiliation taxonomique implémentée dans ces outils est basée sur des approches probabilistes et d'alignement, et l'information concernant la présence de groupes monophylétiques et leurs éventuelles implications écologiques n'est pas restituée. L'outil développé dans ce travail, PANAM, implémente une approche phylogénétique pour affilier les séquences issues des NGS, et nous avons montré que les résultats d'identification taxonomique qu'il produit sont plus précis que les autres outils. De plus, outre les étapes de contrôle de qualité des séquences requêtes, de calcul des indices de richesse et de diversité, et de la comparaison des écosystèmes, PANAM est le seul outil qui permet à ce jour de décrire des clades environnementaux d'intérêt. Cet outil a été déployé sur un cluster afin d'améliorer ses capacités de calcul, et de proposer un service web intégré pour le traitement des données du pyroséquençage. En effet, alors que PANAM a été initialement conçu pour le traitement d'un million de séquences, le développement rapide des plates-formes de séquençage générant de plus en plus de données (jusqu'à 10 millions de séquences par la technique MiSeq), nécessite l'utilisation des structures de calcul distribué. Au-delà des ressources informatiques, nous souhaitons faire évoluer ce service web pour une meilleure caractérisation des espèces environnementales. En effet, depuis la mise en évidence de la biosphère rare, de nombreuses hypothèses concernant son potentiel rôle dans le rétablissement des communautés microbiennes après une perturbation environnementale, ou la diversité génétique qu'elle peut contenir (Sogin et al., 2006), ont été énoncées. Alors que la profondeur de séquençage actuelle associée à une identification phylogénétique permettrait de décrypter cette diversité, son étude à la lumière de paramètres environnementaux pourrait l'associer à des fonctions particulières et fournir plus d'éléments quant à son rôle dans l'écosystème.

Bibliographie

Bibliographie

Mark Achtman and Michael Wagner. Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, 6(6) :431–440, 2008.

Silvia G Acinas, Luisa A Marcelino, Vanja Klepac-Ceraj, and Martin F Polz. Divergence and redundancy of 16s rrna sequences in genomes with multiple rrn operons. *Journal of Bacteriology*, 186(9) :2629–2635, 2004.

Sina M Adl, Alastair GB Simpson, Mark A Farmer, Robert A Andersen, O Roger Anderson, John R Barta, Samuel S Bowser, Guy Brugerolle, Robert A Fensome, Suzanne Fredericq, et al. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *Journal of Eukaryotic Microbiology*, 52(5) :399–451, 2005.

Sina M Adl, Alastair GB Simpson, Christopher E Lane, Julius Lukeš, David Bass, Samuel S Bowser, Matthew W Brown, Fabien Burki, Micah Dunthorn, Vladimir Hampl, et al. The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, 59 (5) :429–514, 2012.

Hélène Agogué, Dominique Lamy, Phillip R Neal, Mitchell L Sogin, and Gerhard J Herndl. Water mass-specificity of bacterial communities in the north atlantic revealed by massively parallel sequencing. *Molecular ecology*, 20(2) :258–274, 2011.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic acids research*, 25(17) :3389–3402, 1997.

Rudolf I Amann, Wolfgang Ludwig, and Karl-Heinz Schleifer. Phylogenetic identification

and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1) :143–169, 1995.

Anthony S Amend, Keith A Seifert, and Thomas D Bruns. Quantifying microbial communities with 454 pyrosequencing : does read abundance count ? *Molecular Ecology*, 19 (24) :5555–5565, 2010.

Anders F Andersson, Mathilda Lindberg, Hedvig Jakobsson, Fredrik Bäckhed, Pål Nyrén, and Lars Engstrand. Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PloS one*, 3(7) :e2836, 2008.

Samuel V Angiuoli, Malcolm Matalka, Aaron Gussman, Kevin Galens, Mahesh Vangala, David R Riley, Cesar Arze, James R White, Owen White, and W Florian Fricke. Clovr : A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC bioinformatics*, 12(1) :356, 2011.

Charles Bachy, John R Dolan, Purificación López-García, Philippe Deschamps, and David Moreira. Accuracy of protist diversity assessments : morphology compared with cloning and direct pyrosequencing of 18s rrna genes and its regions using the conspicuous tintinnid ciliates as a case study. *The ISME Journal*, 7(2) :244–255, 2012.

Anke Behnke, Matthias Engel, Richard Christen, Markus Nebel, Rolf R Klein, and Thorsten Stoeck. Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable ssu rrna gene regions. *Environmental microbiology*, 13 (2) :340–349, 2011.

David Berry, Karim Ben Mahfoudh, Michael Wagner, and Alexander Loy. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and environmental microbiology*, 77(21) :7846–7849, 2011.

Holly M Bik, Way Sung, Paul De Ley, James G Baldwin, Jyotsna Sharma, Axayácatl Rocha-Olivares, and W Kelley Thomas. Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments.

Marc J Bonder, Sanne Abeln, Egija Zaura, and Bernd W Brandt. Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics*, 28(22) :2891–2897, 2012.

Delphine Boucher, Ludwig Jardillier, and Didier Debroas. Succession of bacterial community composition over two consecutive years in two aquatic systems : a natural lake and a lake-reservoir. *FEMS microbiology ecology*, 55(1) :79–97, 2006.

Jennifer L Bowen, Hilary G Morrison, John E Hobbie, and Mitchell L Sogin. Salt marsh sediment diversity : a test of the variability of the rare biosphere among environmental replicates. *The ISME Journal*, 2012.

Vincent Breton, Ana L da Costa, Paul de Vlieger, Young-Min Kim, Lydia Maigne, Romain Reuillon, David Sarramia, Nam Hai Truong, Hong Quang Nguyen, Doman Kim, et al. Innovative in silico approaches to address avian flu using grid technology. *Infectious Disorders-Drug Targets*, 9(3) :358–365, 2009.

Fabien Burki, Kamran Shalchian-Tabrizi, Marianne Minge, Åsmund Skjæveland, Sergey I Nikolaev, Kjetill S Jakobsen, and Jan Pawlowski. Phylogenomics reshuffles the eukaryotic supergroups. *PloS one*, 2(8) :e790, 2007.

Barbara J Campbell, Liying Yu, John F Heidelberg, and David L Kirchman. Activity of abundant and rare bacteria in a coastal ocean. *Proceedings of the National Academy of Sciences*, 108(31) :12776–12781, 2011.

J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5) :335–336, 2010.

J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, et al. Moving pictures of the human microbiome. *Genome Biol*, 12(5) :R50, 2011a.

J Gregory Caporaso, Konrad Paszkiewicz, Dawn Field, Rob Knight, and Jack A Gilbert. The western english channel contains a persistent microbial seed bank. *The ISME journal*, 6(6) :1089–1093, 2011b.

David A Caron and Peter D Countway. Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquatic Microbial Ecology*, 57(3) :227, 2009.

David A Caron, Peter D Countway, Pratik Savai, Rebecca J Gast, Astrid Schnetzer, Stefanie D Moorthi, Mark R Dennett, Dawn M Moran, and Adriane C Jones. Defining dna-based operational taxonomic units for microbial-eukaryote ecology. *Applied and environmental microbiology*, 75(18) :5797–5808, 2009.

Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, pages 265–270, 1984.

Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87(417) :210–217, 1992.

Lu Cheng, Alan W Walker, and Jukka Corander. Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic acids research*, 40(12) :5240–5249, 2012.

Man Kit Cheung, Chun Hang Au, Ka Hou Chu, Hoi Shan Kwan, and Chong Kim Wong. Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. *The ISME journal*, 4(8) :1053–1059, 2010.

Guy Cochrane, Ilene Karsch-Mizrachi, and Yasukazu Nakamura. The international nucleotide sequence database collaboration. *Nucleic acids research*, 39(suppl 1) :D15–D18, 2011.

Frederick M Cohan. Sexual isolation and speciation in bacteria. In *Genetics of Mate Choice : From Sexual Selection to Sexual Isolation*, pages 359–370. Springer, 2002.

James R Cole, Qiong Wang, Benli Chai, and James M Tiedje. The ribosomal database project : Sequences and software for high-throughput rrna analysis. *Handbook of molecular microbial ecology I : Metagenomics and complementary approaches. John Wiley & Sons, Hoboken, NJ*, pages 313–324, 2011.

JR Cole, B Chai, RJ Farris, Q Wang, AS Kulam-Syed-Mohideen, DM McGarrell, AM Bandela, E Cardenas, GM Garrity, and JM Tiedje. The ribosomal database project (rdp-ii) : introducing myrdp space and quality controlled public data. *Nucleic Acids Research*, 35(suppl 1) :D169–D172, 2007.

JR Cole, Q Wang, E Cardenas, J Fish, B Chai, RJ1 Farris, AS Kulam-Syed-Mohideen, DM McGarrell, T Marsh, GM Garrity, et al. The ribosomal database project : improved

alignments and new tools for rrna analysis. *Nucleic acids research*, 37(suppl 1) :D141–D145, 2009.

Robert K Colwell and Jonathan A Coddington. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences*, 345(1311) :101–118, 1994.

André M Comeau, Tommy Harding, Pierre E Galand, Warwick F Vincent, and Connie Lovejoy. Vertical distribution of microbial communities in a perennially stratified arctic lake with saline, anoxic bottom waters. *Scientific reports*, 2, 2012.

Florence Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, 16(22) :10881–10890, 1988.

R Cronn, M Cedroni, T Haselkorn, C Grover, and Jonathan F Wendel. Pcr-mediated recombination in amplification products derived from polyploid cotton. *Theoretical and Applied Genetics*, 104(2-3) :482–489, 2002.

Kara Bowen De León, Bradley D Ramsay, and Matthew W Fields. Quality-score refinement of ssu rrna gene pyrosequencing differs across gene region for environmental samples. *Microbial ecology*, 64(2) :499–508, 2012.

Didier Debroas, Jean-François Humbert, François Enault, Gisèle Bronner, Michael Faubladier, and Emmanuel Cornillot. Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (lac du bourget–france). *Environmental microbiology*, 11(9) :2412–2424, 2009.

Todd Z DeSantis, Carol E Stone, Sonya R Murray, Jordan P Moberg, and Gary L Andersen. Rapid quantification and taxonomic classification of environmental dna from both prokaryotic and eukaryotic origins using a microarray. *FEMS microbiology letters*, 245 (2) :271–278, 2005.

Todd Z DeSantis, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, 72(7) :5069–5072, 2006.

G Devulder, G Perriere, F Baty, and JP Flandrois. Bibi, a bioinformatics bacterial identification tool. *Journal of clinical microbiology*, 41(4) :1785–1787, 2003.

TT Doan. *Epidemiologie moléculaire et métagénomique à haut débit sur la grille*. PhD thesis, Université Blaise Pascal-Clermont-Ferrand II, 2012.

Sean R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9) :755–763, 1998.

Robert C Edgar. Muscle : multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5) :1792–1797, 2004.

Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19) :2460–2461, 2010.

Robert C Edgar, Brian J Haas, Jose C Clemente, Christopher Quince, and Rob Knight. Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16) : 2194–2200, 2011.

Anna Engelbrektson, Victor Kunin, Kelly C Wrighton, Natasha Zvenigorodsky, Feng Chen, Howard Ochman, and Philip Hugenholtz. Experimental factors affecting pcr-based estimates of microbial species richness and evenness. *The ISME journal*, 4(5) : 642–647, 2010.

Paul G Falkowski, Tom Fenchel, and Edward F Delong. The microbial engines that drive earth's biogeochemical cycles. *Science*, 320(5879) :1034–1039, 2008.

Larry M Feinstein, Woo Jun Sul, and Christopher B Blackwood. Assessment of bias associated with incomplete extraction of microbial dna from soil. *Applied and Environmental Microbiology*, 75(16) :5428–5433, 2009.

Tom Fenchel. Cosmopolitan microbes and their cryptic' species. *Aquatic Microbial Ecology*, 41(1) :49–54, 2005.

Noah Fierer, Mya Breitbart, James Nulton, Peter Salamon, Catherine Lozupone, Ryan Jones, Michael Robeson, Robert A Edwards, Ben Felts, Steve Rayhawk, et al. Metagenomic and small-subunit rrna analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and environmental microbiology*, 73(21) :7059–7066, 2007.

Noah Fierer, Jonathan W Leff, Byron J Adams, Uffe N Nielsen, Scott Thomas Bates, Christian L Lauber, Sarah Owens, Jack A Gilbert, Diana H Wall, and J Gregory Caporaso. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, 109(52) :21390–21395, 2012.

Jed A Fuhrman. Microbial community structure and its functional implications. *Nature*, 459(7244) :193–199, 2009.

Jed A Fuhrman and Åke Hagström. Bacterial and archaeal community structure and its patterns. *Microbial Ecology of the Oceans, Second Edition*, pages 45–90, 2008.

Jed A Fuhrman, Ian Hewson, Michael S Schwalbach, Joshua A Steele, Mark V Brown, and Shahid Naeem. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences*, 103(35) :13104–13109, 2006.

Pierre E Galand, Emilio O Casamayor, David L Kirchman, and Connie Lovejoy. Ecology of the rare microbial biosphere of the arctic ocean. *Proceedings of the National Academy of Sciences*, 106(52) :22427–22432, 2009.

Pierre E Galand, Carmen Gutiérrez-Provecho, Ramon Massana, Josep M Gasol, Emilio O Casamayor, et al. Inter-annual recurrence of archaeal assemblages in the coastal nw mediterranean sea (blanes bay microbial observatory). 2010.

Olivier Gascuel. Bionj : an improved version of the nj algorithm based on a simple model of sequence data. *Molecular biology and evolution*, 14(7) :685–695, 1997.

Scott M Gifford, Shalabh Sharma, Johanna M Rinta-Kanto, and Mary Ann Moran. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *The ISME journal*, 5(3) :461–472, 2010.

Thomas M Gihring, Stefan J Green, and Christopher W Schadt. Massively parallel rrna gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environmental microbiology*, 14(2) :285–290, 2012.

André Gilles, Emese Meglécz, Nicolas Pech, Stéphanie Ferreira, Thibaut Malausa, and Jean-François Martin. Accuracy and quality assessment of 454 gs-flx titanium pyrosequencing. *Bmc Genomics*, 12(1) :245, 2011.

Adriana Giongo, David B Crabb, Austin G Davis-Richardson, Diane Chauliac, Jennifer M Mobberley, Kelsey A Gano, Nabanita Mukherjee, George Casella, Luiz FW Roesch, Brandon Walts, et al. Pangea : pipeline for analysis of next generation amplicons. *The ISME journal*, 4(7) :852–861, 2010.

Travis C Glenn. Field guide to next-generation dna sequencers. *Molecular Ecology Resources*, 11(5) :759–769, 2011.

Angélique Gobet, Christopher Quince, and Alban Ramette. Multivariate cutoff level analysis (multicola) of large community data sets. *Nucleic acids research*, 38(15) : e155–e155, 2010.

Brett M Goebel and Erko Stackebrandt. Cultural and phylogenetic analysis of mixed microbial populations found in natural and commercial bioleaching environments. *Applied and Environmental Microbiology*, 60(5) :1614–1621, 1994.

Jeremy Goecks, Anton Nekrutenko, James Taylor, T Galaxy Team, et al. Galaxy : a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8) :R86, 2010.

Jun Gong, Jun Dong, Xihan Liu, and Ramon Massana. Extremely high copy numbers and polymorphisms of the rdna operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist*, 2013.

Antonio Gonzalez and Rob Knight. Advancing analytical algorithms and pipelines for billions of microbial sequences. *Current Opinion in Biotechnology*, 23(1) :64–71, 2012.

Dan Graur and Wen-Hsiung Li. *Fundamentals of molecular evolution*, volume 2. Sinauer Associates Sunderland, MA, 2000.

Peter M Groffman, Mark A Altabet, JK Böhlke, Klaus Butterbach-Bahl, Mark B David, Mary K Firestone, Anne E Giblin, Todd M Kana, Lars Peter Nielsen, and Mary A Voytek. Methods for measuring denitrification : diverse approaches to a difficult problem. *Ecological Applications*, 16(6) :2091–2122, 2006.

Laure Guillou, Dipankar Bachar, Stéphane Audic, David Bass, Cédric Berney, Lucie Bittner, Christophe Boutte, Gaétan Burgaud, Colomban De Vargas, Johan Decelle, et al. The protist ribosomal reference database (pr2) : a catalog of unicellular eukaryote small sub-unit rrna sequences with curated taxonomy. *Nucleic acids research*, 41(D1) :D597–D604, 2013.

Stéphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5) :696–704, 2003.

Brian J Haas, Dirk Gevers, Ashlee M Earl, Mike Feldgarden, Doyle V Ward, Georgia Giannoukos, Dawn Ciulla, Diana Tabbaa, Sarah K Highlander, Erica Sodergren, et al. Chimeric 16s rrna sequence formation and detection in sanger and 454-pyrosequenced pcr amplicons. *Genome research*, 21(3) :494–504, 2011.

Mehrdad Hajibabaei, Shadi Shokralla, Xin Zhou, Gregory AC Singer, and Donald J Baird. Environmental barcoding : a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, 6(4) :e17497, 2011.

Micah Hamady and Rob Knight. Microbial community profiling for human microbiome projects : Tools, techniques, and challenges. *Genome research*, 19(7) :1141–1152, 2009.

Xiaolin Hao, Rui Jiang, and Ting Chen. Clustering 16s rrna for otu prediction : a method of unsupervised bayesian clustering. *Bioinformatics*, 27(5) :611–618, 2011.

J Kirk Harris, J Gregory Caporaso, Jeffrey J Walker, John R Spear, Nicholas J Gold, Charles E Robertson, Philip Hugenholtz, Julia Goodrich, Daniel McDonald, Dan Knights, et al. Phylogenetic stratigraphy in the guerrero negro hypersaline microbial mat. *The ISME journal*, 2012.

Amber Hartman, Sean Riddle, Timothy McPhillips, Bertram Ludäscher, and Jonathan Eisen. Introducing waters : a workflow for the alignment, taxonomy, and ecology of ribosomal sequences. *BMC bioinformatics*, 11(1) :317, 2010.

Julie A Huber, David B Mark Welch, Hilary G Morrison, Susan M Huse, Phillip R Neal, David A Butterfield, and Mitchell L Sogin. Microbial population structures in the deep marine biosphere. *science*, 318(5847) :97–100, 2007.

Jennifer B Hughes, Jessica J Hellmann, Taylor H Ricketts, and Brendan JM Bohannan. Counting the uncountable : statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, 67(10) :4399–4406, 2001.

Susan M Huse, Julie A Huber, Hilary G Morrison, Mitchell L Sogin, David Mark Welch, et al. Accuracy and quality of massively parallel dna pyrosequencing. *Genome biol*, 8 (7) :R143, 2007.

Susan M Huse, Les Dethlefsen, Julie A Huber, David Mark Welch, David A Relman, and Mitchell L Sogin. Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing. *PLoS genetics*, 4(11) :e1000255, 2008.

Susan M Huse, David Mark Welch, Hilary G Morrison, and Mitchell L Sogin. Ironing out the wrinkles in the rare biosphere through improved otu clustering. *Environmental microbiology*, 12(7) :1889–1898, 2010.

Patricio Jeraldo, Nicholas Chia, and Nigel Goldenfeld. On the suitability of short reads of 16s rrna for phylogeny-based analyses in environmental surveys. *Environmental microbiology*, 13(11) :3000–3009, 2011.

Stuart E Jones and Jay T Lennon. Dormancy contributes to the maintenance of microbial diversity. *Proceedings of the National Academy of Sciences*, 107(13) :5881–5886, 2010.

Ari Jumpponen. Soil fungal communities underneath willow canopies on a primary successional glacier forefront : rdna sequence results can be affected by primer selection and chimeric data. *Microbial Ecology*, 53(2) :233–246, 2007.

Paul F Kemp and Josephine Y Aller. Bacterial diversity in aquatic and other environments : what 16s rdna libraries can tell us. *FEMS Microbiology Ecology*, 47(2) :161–177, 2004.

Minseok Kim, Mark Morrison, and Zhongtang Yu. Evaluation of different partial 16s rrna gene sequence regions for phylogenetic analysis of microbiomes. *Journal of microbiological methods*, 84(1) :81–87, 2011.

David L Kirchman, Matthew T Cottrell, and Connie Lovejoy. The structure of bacterial communities in the western arctic ocean as revealed by pyrosequencing of 16s rrna genes. *Environmental microbiology*, 12(5) :1132–1143, 2010.

Alexander F Koeppel and Martin Wu. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial operational taxonomic units. *Nucleic acids research*, 2013.

Allan Konopka. Microbial ecology : searching for principles. *Microbe*, 1(4) :175–179, 2006.

V Kunin and P Hugenholtz. Pyrotagger : A fast, accurate pipeline for analysis of rrna amplicon pyrosequence data. *The Open Journal*, 1(1), 2010.

Victor Kunin, Anna Engelbrektson, Howard Ochman, and Philip Hugenholtz. Wrinkles in the rare biosphere : pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental microbiology*, 12(1) :118–123, 2010.

MA Larkin, G Blackshields, NP Brown, R Chenna, PA McGettigan, H McWilliam, F Valentin, IM Wallace, A Wilm, R Lopez, et al. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21) :2947–2948, 2007.

Jeffrey G Lawrence. Gene transfer, speciation, and the evolution of bacterial genomes. *Current opinion in microbiology*, 2(5) :519–523, 1999.

Charles K Lee, Craig W Herbold, Shawn W Polson, K Eric Wommack, Shannon J Williamson, Ian R McDonald, and S Craig Cary. Groundtruthing next-gen sequencing for microbial ecology–biases and errors in community structure estimates from pcr amplicon pyrosequencing. *PLoS One*, 7(9) :e44224, 2012.

Zarraz May-Ping Lee, Carl Bussema, and Thomas M Schmidt. rrndb : documenting the number of rrna and trna genes in bacteria and archaea. *Nucleic Acids Research*, 37 (suppl 1) :D489–D493, 2009.

Marie Lefranc, Aurélie Thénot, Cécile Lepere, and Didier Debroas. Genetic diversity of small eukaryotes in lakes differing by their trophic status. *Applied and environmental microbiology*, 71(10) :5935–5942, 2005.

Frederik Leliaert, David R Smith, Hervé Moreau, Matthew D Herron, Heroen Verbruggen, Charles F Delwiche, and Olivier De Clerck. Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences*, 31(1) :1–46, 2012.

Leandro N Lemos, Roberta R Fulthorpe, Eric W Triplett, and Luiz FW Roesch. Rethinking microbial diversity analysis in the high throughput sequencing era. *Journal of microbiological methods*, 86(1) :42–51, 2011.

Ivica Letunic and Peer Bork. Interactive tree of life (itol) : an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1) :127–128, 2007.

Ivica Letunic and Peer Bork. Interactive tree of life v2 : online annotation and display of phylogenetic trees made easy. *Nucleic acids research*, 39(suppl 2) :W475–W478, 2011.

Weizhong Li, Lukasz Jaroszewski, and Adam Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3) :282–283, 2001.

Kevin Liu, C Randal Linder, and Tandy Warnow. Raxml and fasttree : comparing two methods for large-scale maximum likelihood phylogeny estimation. *PloS one*, 6(11) : e27731, 2011.

Zongzhi Liu, Catherine Lozupone, Micah Hamady, Frederic D Bushman, and Rob Knight. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic acids research*, 35(18) :e120, 2007.

Zongzhi Liu, Todd Z DeSantis, Gary L Andersen, and Rob Knight. Accurate taxonomy assignments from 16s rrna sequences produced by highly parallel pyrosequencers. *Nucleic acids research*, 36(18) :e120–e120, 2008.

Ramiro Logares, Thomas HA Haverkamp, Surendra Kumar, Anders Lanzén, Alexander J Nederbragt, Christopher Quince, and Håvard Kauserud. Environmental microbiology through the lens of high-throughput dna sequencing : Synopsis of current platforms and bioinformatics approaches. *Journal of microbiological methods*, 2012.

Nicholas J Loman, Raju V Misra, Timothy J Dallman, Chrystala Constantinidou, Saheer E Gharbia, John Wain, and Mark J Pallen. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5) :434–439, 2012.

Catherine Lozupone and Rob Knight. Unifrac : a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12) :8228–8235, 2005.

Catherine Lozupone, Micah Hamady, and Rob Knight. Unifrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC bioinformatics*, 7(1) :371, 2006.

John A Ludwig and James F Reynolds. *Statistical ecology : a primer in methods and computing*. Wiley-Interscience, 1988.

Wolfgang Ludwig, Oliver Strunk, Ralf Westram, Lothar Richter, Harald Meier, Arno Buchner, Tina Lai, Susanne Steppi, Gangolf Jobb, Wolfram Förster, et al. Arb : a software environment for sequence data. *Nucleic acids research*, 32(4) :1363–1371, 2004.

Daniel Lundin, Ina Severin, Jürg Brendan Logue, Örjan Östman, Anders F Andersson, and Eva S Lindström. Which sequencing depth is sufficient to describe patterns in bacterial α-and β-diversity ? *Environmental Microbiology Reports*, 4(3) :367–372, 2012.

Birger Marin and Michael Melkonian. Molecular phylogeny and classification of the mamiellophyceae class. nov.(chlorophyta) based on sequence comparisons of the nuclear- and plastid-encoded rrna operons. *Protist*, 161(2) :304–336, 2010.

Andrew P Martin. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and environmental microbiology*, 68(8) :3673–3682, 2002.

Frederick Matsen, Robin Kodner, and E Virginia Armbrust. pplacer : linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11(1) :538, 2010.

Ernst Mayr. *Systematics and the origin of species, from the viewpoint of a zoologist*. Number 13. Harvard University Press, 1942.

Ralph Medinger, Viola Nolte, Ram Vinay Pandey, Steffen Jost, Birgit Ottenwälder, Christian Schlötterer, and Jens Boenigk. Diversity in a hidden world : potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology*, 19(s1) :32–40, 2010.

Sergio E Morales, Theodore F Cosart, Jesse V Johnson, and William E Holben. Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness

and the effects of amplicon length, degree of coverage, and dna fractionation on classification and ecological parameters. *Applied and environmental microbiology*, 75(3) : 668–675, 2009.

Viola Nolte, Ram Vinay Pandey, Steffen Jost, Ralph Medinger, Birgit Ottenwaelder, Jens Boenigk, and Christian Schlotterer. Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Molecular ecology*, 19(14) : 2908–2915, 2010.

Ron O'Dor, Patricia Miloslavich, and Kristen Yarincik. Marine biodiversity and biogeography–regional comparisons of global issues, an introduction. *PloS one*, 5(8) : e11871, 2010.

Norman R Pace. A molecular view of microbial diversity and the biosphere. *Science*, 276 (5313) :734–740, 1997.

Ram V Pandey, Viola Nolte, and Christian Schlötterer. Cangs : a user-friendly utility for processing and analyzing 454 gs-flx data in biodiversity studies. *BMC research notes*, 3(1) :3, 2010.

David J Patterson. Seeing the big picture on microbe distribution. *Science*, 325(5947) : 1506–1507, 2009.

Jan Pawlowski, Richard Christen, Béatrice Lecroq, Dipankar Bachar, Hamid Reza Shahbazkia, Linda Amaral-Zettler, and Laure Guillou. Eukaryotic richness in the abyss : insights from pyrotag sequencing. *PLoS One*, 6(4) :e18169, 2011.

Carlos Pedrós-Alió. Marine microbial diversity : can it be determined ? *Trends in microbiology*, 14(6) :257–263, 2006.

Martin F Polz and Colleen M Cavanaugh. Bias in template-to-product ratios in multitemplate pcr. *Applied and Environmental Microbiology*, 64(10) :3724–3730, 1998.

Jeff R Powell, Michael T Monaghan, Maarja Öpik, and Matthias C Rillig. Evolutionary criteria outperform operational approaches in producing ecologically relevant fungal species inventories. *Molecular Ecology*, 20(3) :655–666, 2011.

Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree : computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7) :1641–1650, 2009.

Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *Plos one*, 5(3) :e9490, 2010.

Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. Silva : a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids research*, 35(21) :7188–7196, 2007.

Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418) :55–60, 2012.

Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project : improved data processing and web-based tools. *Nucleic acids research*, 41(D1) :D590–D596, 2013.

Christopher Quince, Anders Lanzén, Thomas P Curtis, Russell J Davenport, Neil Hall, Ian M Head, L Fiona Read, and William T Sloan. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature methods*, 6(9) :639–641, 2009.

Christopher Quince, Anders Lanzen, Russell J Davenport, and Peter J Turnbaugh. Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, 12(1) :38, 2011.

Aaron R Quinlan, Donald A Stewart, Michael P Strömberg, and Gábor T Marth. Pyrobayes : an improved base caller for snp discovery in pyrosequences. *Nature methods*, 5 (2) :179–181, 2008.

Benjamin Ragan-Kelley, William Anton Walters, Daniel McDonald, Justin Riley, Brian E Granger, Antonio Gonzalez, Rob Knight, Fernando Perez, and J Gregory Caporaso. Collaborative cloud-enabled tools allow rapid, reproducible biological insights. *The ISME journal*, 2012.

Jens Reeder and Rob Knight. Rapid denoising of pyrosequencing amplicon data : exploiting the rank-abundance distribution. *Nature methods*, 7(9) :668, 2010.

Luiz FW Roesch, Roberta R Fulthorpe, Alberto Riva, George Casella, Alison KM Hadwin, Angela D Kent, Samira H Daroub, Flavio AO Camargo, William G Farmerie, and Eric W Triplett. Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, 1(4) :283–290, 2007.

Alejandro P Rooney and Todd J Ward. Evolution of a large ribosomal rna multigene family in filamentous fungi : birth and death of a concerted evolution paradigm. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14) : 5084–5089, 2005.

Gabriella Rozera, Isabella Abbate, Alessandro Bruselles, Crhysoula Vlassi, Gianpiero D'Offizi, Pasquale Narciso, Giovanni Chillemi, Mattia Prosperi, Giuseppe Ippolito, and Maria Capobianchi. Massively parallel pyrosequencing highlights minority variants in the hiv-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology*, 6(1) :15, 2009.

Patrick D Schloss and Jo Handelsman. Introducing dotur, a computer program for defining operational taxonomic units and estimating species richness. *Applied and environmental microbiology*, 71(3) :1501–1506, 2005.

Patrick D Schloss and Sarah L Westcott. Assessing and improving methods used in operational taxonomic unit-based approaches for 16s rrna gene sequence analysis. *Applied and environmental microbiology*, 77(10) :3219–3226, 2011.

Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur : open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23) :7537–7541, 2009.

Heiko A Schmidt, Korbinian Strimmer, Martin Vingron, and Arndt von Haeseler. Tree-puzzle : maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3) :502–504, 2002.

Matthew B Scholz, Chien-Chi Lo, and Patrick SG Chain. Next generation sequencing and bioinformatic bottlenecks : the current state of metagenomic data analysis. *Current opinion in biotechnology*, 23(1) :9–15, 2012.

Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26 (10) :1135–1145, 2008.

Hidetoshi Shimodaira and Masami Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular biology and evolution*, 16 :1114–1116, 1999.

Shadi Shokralla, Jennifer L Spall, Joel F Gibson, and Mehrdad Hajibabaei. Next-generation sequencing technologies for environmental dna research. *Molecular Ecology*, 21(8) :1794–1805, 2012.

Edward H Simpson. Measurement of diversity. *Nature*, 163(4148) :688, 1949.

Mitchell L Sogin, Hilary G Morrison, Julie A Huber, David Mark Welch, Susan M Huse, Phillip R Neal, Jesus M Arrieta, and Gerhard J Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, 103(32) :12115–12120, 2006.

E Stackebrandt and BM Goebel. Taxonomic note : a place for dna-dna reassociation and 16s rrna sequence analysis in the present species definition in bacteriology. *International Journal of Systematic Bacteriology*, 44(4) :846–849, 1994.

Alexandros Stamatakis. Raxml-vi-hpc : maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21) :2688–2690, 2006.

Alexandros Stamatakis, Paul Hoover, and Jacques Rougemont. A rapid bootstrap algorithm for the raxml web servers. *Systematic biology*, 57(5) :758–771, 2008.

Lincoln D Stein et al. The case for cloud computing in genome informatics. *Genome Biol*, 11(5) :207, 2010.

Thorsten Stoeck, David Bass, Markus Nebel, Richard Christen, Meredith DM Jones, HANS-WERNER BREINER, and Thomas A Richards. Multiple marker parallel tag

environmental dna sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, 19(s1) :21–31, 2010.

Yijun Sun, Yunpeng Cai, Li Liu, Fahong Yu, Michael L Farrell, William McKendree, and William Farmerie. Esprit : estimating species richness using large collections of 16s rrna pyrosequences. *Nucleic acids research*, 37(10) :e76–e76, 2009.

Yijun Sun, Yunpeng Cai, Susan M Huse, Rob Knight, William G Farmerie, Xiaoyu Wang, and Volker Mai. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in bioinformatics*, 13(1) :107–121, 2012.

Julie D Thompson, Desmond G Higgins, and Toby J Gibson. Clustal w : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22 (22) :4673–4680, 1994.

Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164) :804–810, 2007.

Peter J Turnbaugh, Fredrik Bäckhed, Lucinda Fulton, Jeffrey I Gordon, et al. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell host & microbe*, 3(4) :213, 2008a.

Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228) :480–484, 2008b.

Friedrich V Wintzingerode, Ulf B Göbel, and Erko Stackebrandt. Determination of microbial diversity in environmental samples : pitfalls of pcr-based rrna analysis. *FEMS microbiology reviews*, 21(3) :213–229, 1997.

Alice Valentini, Christian Miquel, MUHAMMAD ALI NAWAZ, EVA Bellemain, Eric Coissac, FranCOis Pompanon, Ludovic Gielly, Corinne Cruaud, Giuseppe Nascetti, Patrick Wincker, et al. New perspectives in diet analysis based on dna barcoding and

parallel pyrosequencing : the trnl approach. *Molecular Ecology Resources*, 9(1) :51–60, 2009.

J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, et al. Environmental genome shotgun sequencing of the sargasso sea. *science*, 304(5667) : 66–74, 2004.

Kevin L Vergin, Bánk Beszteri, Adam Monier, J Cameron Thrash, Ben Temperton, Alexander H Treusch, Fabian Kilpert, Alexandra Z Worden, and Stephen J Giovannoni. High-resolution sar11 ecotype dynamics at the bermuda atlantic time-series study site by phylogenetic placement of pyrosequences. *The ISME Journal*, 2013.

Patrick Vigil, Peter D Countway, Julie Rose, Darcy J Lonsdale, Christopher J Gobler, David A Caron, et al. Rapid shifts in dominant taxa among microbial eukaryotes in estuarine ecosystems. *Aquatic Microbial Ecology*, 54(1) :83, 2009.

Maria Vila-Costa, Josep M Gasol, Shalabh Sharma, and Mary Ann Moran. Community analysis of high-and low-nucleic acid-containing bacteria in nw mediterranean coastal waters using 16s rdna pyrosequencing. *Environmental microbiology*, 14(6) :1390–1402, 2012.

Maria Vila-Costa, Albert Barberan, Jean-Christophe Auguet, Shalabh Sharma, Mary Ann Moran, and Emilio O Casamayor. Bacterial and archaeal community structure in the surface microlayer of high mountain lakes examined under two atmospheric aerosol loading scenarios. *FEMS microbiology ecology*, 2013.

Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16) :5261–5267, 2007.

David M Ward, Fred M Cohan, Devaki Bhaya, John F Heidelberg, Michael Kühl, and Arthur Grossman. Genomics, environmental genomics and the issue of microbial species. *Heredity*, 100(2) :207–219, 2007.

Jeffrey J Werner, Omry Koren, Philip Hugenholtz, Todd Z DeSantis, William A Walters, J Gregory Caporaso, Largus T Angenent, Rob Knight, and Ruth E Ley. Impact of

training sets on classification of high-throughput bacterial 16s rrna gene surveys. *The ISME journal*, 6(1) :94–103, 2011.

James R White, Saket Navlakha, Niranjan Nagarajan, Mohammad-Reza Ghodsi, Carl Kingsford, and Mihai Pop. Alignment and clustering of phylogenetic markers-implications for microbial diversity studies. *BMC bioinformatics*, 11(1) :152, 2010.

Dongying Wu, Amber Hartman, Naomi Ward, and Jonathan A Eisen. An automated phylogenetic tree-based small subunit rrna taxonomy and alignment pipeline (stap). *PloS one*, 3(7) :e2566, 2008.

Noha Youssef, Cody S Sheik, Lee R Krumholz, Fares Z Najar, Bruce A Roe, and Mostafa S Elshahed. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16s rrna gene-based environmental surveys. *Applied and environmental microbiology*, 75(16) :5227–5236, 2009.

Egija Zaura, Bart Keijser, Susan Huse, and Wim Crielaard. Defining the healthy. *BMC microbiology*, 9(1) :259, 2009.

Jizhong Zhou, Liyou Wu, Ye Deng, Xiaoyang Zhi, Yi-Huei Jiang, Qichao Tu, Jianping Xie, Joy D Van Nostrand, Zhili He, and Yunfeng Yang. Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME journal*, 5(8) :1303–1313, 2011.

Annexe

RESEARCH ARTICLE

Geographic distance and ecosystem size determine the distribution of smallest protists in lacustrine ecosystems

Cécile Lepère[1,2,3], Isabelle Domaizon[3], Najwa Taïb[1,2], Jean-François Mangot[1,2,3], Gisèle Bronner[1,2], Delphine Boucher[1,2,] & Didier Debroas [1,2]

[1]Laboratoire "Microorganismes: Génome et Environnement", Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France; [2]CNRS, UMR 6023, LMGE, Aubière, France and [3]INRA, UMR 42 CARRTEL, Thonon les bains, France

Correspondence: Didier Debroas, Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes : Génome et Environnement", BP 10448, F-63000 Clermont-Ferrand, France.
Tel.: +33 473407837; fax: +33 473407670; e-mail: didier.debroas@univ-bpclermont.fr

Present address: Delphine Boucher, EA 4678 Conception, Ingénierie et Développement de l'Aliment et du Médicament, Clermont Université, Université d'Auvergne, BP 38, 63001, Clermont-Ferrand, France

Received 28 November 2012; revised 18 February 2013; accepted 18 February 2013.

DOI: 10.1111/1574-6941.12100

Editor: Riks Laanbroek

Keywords
biogeography; protists; alpha-diversity; beta-diversity; lakes.

Abstract

Understanding the spatial distribution of aquatic microbial diversity and the underlying mechanisms causing differences in community composition is a challenging and central goal for ecologists. Recent insights into protistan diversity and ecology are increasing the debate over their spatial distribution. In this study, we investigate the importance of spatial and environmental factors in shaping the small protists community structure in lakes. We analyzed small protists community composition (beta-diversity) and richness (alpha-diversity) at regional scale by different molecular methods targeting the gene coding for 18S rRNA gene (T-RFLP and 454 pyrosequencing). Our results show a distance–decay pattern for rare and dominant taxa and the spatial distribution of the latter followed the prediction of the island biogeography theory. Furthermore, geographic distances between lakes seem to be the main force shaping the protists community composition in the lakes studied here. Finally, the spatial distribution of protists was discussed at the global scale (11 worldwide distributed lakes) by comparing these results with those present in the public database. UniFrac analysis showed 18S rRNA gene OTUs compositions significantly different among most of lakes, and this difference does not seem to be related to the trophic status.

Introduction

A main objective in ecology is to understand the dynamic of biodiversity and, more particularly, the spatial repartition of species. The existence of biogeographic patterns has been highlighted in microbial ecology mainly by distance–decay or taxa–area relationships (Green & Bohannan, 2006). Spatial distribution of microbial eukaryotes (i.e. protists) has received much less attention than bacteria and the studies have concentrated mainly on a single taxonomic group. However, their total diversity and distribution in nature are currently the focus of active debates (Finlay & Fenchel, 2004; Foissner, 2006; Pinel-Alloul & Ghadouani, 2007; Caron, 2009; Nolte *et al.*, 2010), and no consensus has been reached yet. Opposing views have asserted that this highly diverse group of organisms presents a cosmopolitan distribution on the one hand (Finlay & Fenchel, 2004; Finlay *et al.*, 2006; Pither, 2007), especially microorganisms smaller than 20 μm according to Yang *et al.* (2010), or is composed primarily of species that have limited geographic distributions on the other (Papke & Ward, 2004; Telford *et al.*, 2006). The biogeography studies about microorganisms seem therefore varied according to the group targeted but also to the methods used for delineating 'microbial species' or/and spatial scale (Lindstrom & Langenheder, 2012). The importance of the methods used was, for instance, underlined by Taylor *et al.* (2006) in a study on fungal strains that showed a global distribution by morphological inspection and endemicity by phylogenetic recognition. In addition, most of the studies might have undersampled microbial communities, and novel high-throughput sequencing technologies might provide

FEMS MICROBIOLOGY ECOLOGY

the tools necessary to explore microbial diversity to a greater depth (Pedros-Alio, 2006; Sogin *et al.*, 2006). Microbial community surveys likely underestimated the slope of the taxa–area relationship, and the evolution of the abundance of the rarest taxa with increasing area was not detectable. Finally, some works omitted to take into account environmental parameters (see review Martiny *et al.*, 2006), and there was no clear distinction between the spatial variation due to present-day environmental factors or historical contingencies.

As model organism to understand the spatial repartition of microbial species in nature, we studied small protists (0.2–5 μm; i.e. unicellular eukaryotic organisms from algae to heterotrophic flagellates, including unicellular fungi (Caron, 2009)) and their distribution in lakes. As well as for bacteria, these small protists are characterized by a tiny size and are likely to disperse easily (ex: capacity of survival during transport for dormant cells) and could have a cosmopolitan distribution, leading to the classical dictum 'everything is everywhere, but the environment selects' (Baas-Becking, 1934). However, protists constitute complex assemblages, and one can wonder if the concept of biogeographic diversity is applicable to small-sized heterogeneous group, which is diverse in terms of physiologies, life cycles, phylogenetic positions, with the ability to reproduce sexually and with capacities of dispersal–colonization, which are likely not the same (i.e. cyst, endospores easily transported in atmosphere/aerosols).

The recent application of molecular approaches to assess the diversity of natural microbial assemblages, mainly in marine environments, has revealed an unexpected diversity, undescribed taxa, and new lineages among these small protists (e.g. Lopez-Garcia *et al.*, 2001) emphasizing they could be as diversified as bacteria. A few recent studies conducted in lakes have also highlighted the wide diversity of 18S rRNA gene sequences affiliated to numerous phylogenetic groups involved in photosynthetic and heterotrophic processes but also in parasitism (e.g. Lefranc *et al.*, 2005; Richards *et al.*, 2005; Lepère *et al.*, 2007, 2008). The application of novel high-throughput sequencing technologies (e.g. 454 pyrosequencing) has revealed a greater diversity within microbial communities (e.g. Sogin *et al.*, 2006) and allows the access to the rarest phylotypes. However, these methods have not commonly been used for microbial eukaryotes especially in lacustrine ecosystems (e.g. Monchy *et al.*, 2011).

In the present study, small protists composition and richness were assessed at a regional scale (six French lakes), by different molecular methods targeting the gene coding for 18S rRNA gene: T-RFLP and 454 pyrosequencing. We tested how change (1) protists richness (alpha-diversity) according to lakes areas and (2) similarity in protists community composition (beta-diversity) according to the geographic distance between lakes. According to Martiny *et al.* (2006), we expected that environmental factors (physical, chemical, and biological local parameters) tested have significant effects in shaping microbial composition at regional scale, whereas distance (dispersal limitation) should be the main structuring factor across continents.

Materials and methods

Studied lakes

The study was conducted in six lakes located in two regions of France (Massif Central and Alps) described in Table 1 (Lakes Godivelle, Aydat, Pavin, and Bourget; Sep and Villerest reservoirs). These lakes were on average 133 km separated from each other, and the most distant lakes are separated by 400 km (Aydat-Bourget). Samples from the six lakes were taken monthly during the thermal stratification period from April to August (from 2002 and 2005). Water samples from the epilimnion (1–5 m) were collected with a Van Dorn bottle at a permanent station situated at the deepest zone of the water column. Water samples [from 100 to 120 mL (maximum volume that can be filtered without clogging the filters)] were successively filtered through 5-μm (as prefiltration step) and 0.2-μm-pore-size polycarbonate filters (Millipore) and stored at −80 °C until nucleic acid extraction. The T-RFLP, a fingerprinting method, was used to examine

Table 1. Main characteristics of the lakes sampled in this study

Lakes	Trophic status	Coordinates	Area (km^2)	Vol (km^3)	Max depth (m)	Mean depth (m)	Altitude (m)
Godivelle	Ultra-oligotrophic	45°23′04″N, 2°55′25″E	0.138	0.003	44	*	1239
Pavin	Oligomesotrophic	45°29′45″N, 2°53′18″E	0.44	0.002	98	54.9	1197
Sep	Oligomesotrophic	46°02′51″N, 3°02′47″E	0.33	0.005	37.4	14.2	414
Bourget	Mesotrophic	45°43′55″N, 5°52′06″E	44.5	3.6	145	81	231.5
Aydat	Eutrophic	45°39′50″N, 2°59′04″ E	0.65	0.004	15.5	7.4	825
Villerest	Hypereutrophic	45°59′36″N, 4°2′12″E	7	0.062	45	18	257

*Unknown.

shifts in the structure of protists community at each site during 4 months (thermal stratification period), whereas pyrosequencing was used to analyze in depth one sample from each of these ecosystems (six lakes). Samples for determining microorganisms' abundances were collected and fixed immediately as described in Lepère *et al.* (2007). Environmental parameters measured in the six lakes, listed in Table S1, were measured as described in Lepère *et al.* (2006).

Molecular methods

Extraction

Nucleic acid extraction has been carried out as described in Lefranc *et al.* (2005), and extracts were stored at −20 °C until analysis.

T-RFLP analysis

PCR, enzymatic digestions (*Msp*I and *Rsa*I), and terminal restriction fragments (T-RFs) analyses were performed as described in Lepère *et al.* (2006). Briefly, samples were analyzed in triplicate, and a T-RF (size between 48 and 560 bp with a peak area > 50 fluorescence) was included in the analysis if it occurred in at least two profiles. To account for small differences in the running time among samples, we considered fragments from different profiles with < 1-base pair difference to be the same length. A program in Visual Basic for Excel was developed to automate these procedures and validated previously (e.g. Lepère *et al.*, 2006). Regardless of lakes, *Msp*I was to be more discriminative enzyme in terms of richness; we have therefore only presented data obtained with *Msp*I. The results were then expressed either in terms of presence or absence, or as a relative percentage area compared with the total area.

Pyrosequencing

The V4-V5 hypervariable regions of eukaryotic 18S rRNA gene were amplified with Ek-NSF573 (5′-CGCGGTAATT CCAGCTCCA-3′) and EK-NSR1147 (5′-CCGTCAATTYY TTTRAGTTT-3′) (Wuyts *et al.*, 2004). To discriminate each sample, a 5-bp multiplex tag was coupled with adaptor A. The amplification mix contained 30 ng of genomic DNA, 200 μM of deoxynucleoside triphosphates (Bioline, London, UK), 2 mM $MgCl_2$ (Bioline), 10 pmol of each primer, 1.5 U of *Taq* DNA polymerase (Bioline), and the PCR buffer. The cycling conditions were an initial denaturation at 94 °C for 10 min followed by 30 cycles of 94 °C for 1 min, 57 °C for 1 min, 72 °C for 1 min and 30 s, and a final 10-min extension at 72 °C. The pyrosequencing data representing 131 869 raw sequence reads were cleaned and analyzed by the method described in (Data S1). Finally, after stringent quality filtering and the subtraction of nonsmall protists affiliated with metazoan and Streptophyta taxa OTUs (operational taxonomic units), a total of 89 337 sequences averaging $\geq$ 200 bp in length were selected for studying small protists (raw data have been deposited in Dryad: http://datadryad.org). The pyrosequencing reads were clustered with a threshold of 95%. The cut-off used for defining an OTU is always debatable; however, the threshold of 95% has been proven to be appropriate to approximate species-level distinction (Caron, 2009; Mangot *et al.*, 2013). The OTUs were compared against our reference database (details in Data S1) with USEARCH (Edgar, 2010), and following the taxonomy of its best hit, each sequence is appended to a phyletic group, together with its 5 best hits. Homologous reads have been then assigned to phyletic groups; they were aligned with the referenced sequences of the corresponding profile using HMMalign (Eddy, 1998). FASTTREE (Price *et al.*, 2009) was used to build phylogenetic trees for each phyletic profile with the Jukes-Cantor + Cat model and a bootstrap threshold of 100 (the trees have been deposited in Dryad: http://datadryad.org).

Data analyses

Lakes can be considered as islands within a 'sea' of land (Dodson, 1992). Therefore, to describe the inter-lake diversity distribution of protists, we tested the 'island biogeography' theory (MacArthur & Wilson, 1967) that proposes that size and distance of an island (lake) determine its richness/diversity (Reche *et al.*, 2005; Smith *et al.*, 2005; Logue *et al.*, 2012). In this study, Margalef index (Hill *et al.*, 2003) that allows computing richness for rare and dominant taxa was used for quantifying the richness (alpha-diversity). The taxa–area relationship (TAR) (Green & Bohannan, 2006) was calculated using: $S \propto A^z$ where S is the richness and A is the area (Table 1). The slope, z, was calculated after log transformation of data. For assessing the independent effect of the factors tested, partial Spearman's rank correlation analyses were performed. This test allows to measure the degree of association between two variables (i.e. richness *vs* area), while a third variable is controlled (i.e. environmental factors). Environmental factors were used as a single variable by introducing in the partial correlation, the first axis data obtained from a principal component analysis (PCA) computed with NH_4-N, NO_3-N, PO_4-P temperature, water clarity (Secchi disk), chlorophyll *a*, prokaryotes, heterotrophic nanoflagellates (HNF), and zooplankton (cladocerans, copepods, and rotifers) abundances (Table S1).

To test the effects of geographic distances vs. environment on assemblage composition (beta-diversity), Mantel (Mantel, 1967) and partial Mantel tests have been completed (Martiny *et al.*, 2006) on the T-RFLP data from the time series and from pyrosequencing data. Although the definition of dominant and rare taxa can be different according to the method used, we defined rare taxa as < 1% (of total area or reads) according to Pedros-Alio (2006). Correlations were carried out between Bray–Curtis community dissimilarity matrix (24 × 24 for T-RFLP and 6 × 6 for pyrosequencing data), a matrix of pairwise lake environmental Euclidian distances (included in Table S1), and a matrix of spatial proximity (distance in kilometers between pairs of lakes). The beta-diversity of protists was also assessed from the UniFrac distance (Lozupone & Knight, 2005) calculated from phylogenies obtained from the pyrosequencing approach (Data S1). We have computed the coverage index ($C_{Good's}$ (Good, 1953) for the pyrosequencing data for each sample, and this index varied between 96.7% (Lake Bourget) and 99.2% (Lake Pavin). These high values demonstrate good diversity coverage.

Results

Composition of the small protist community

On the six studied lakes, on average, eight T-RFs represent more than 60% of the total area, and rare T-RFs account on average for 90% of the total T-RFs number (Fig. S1). The pyrosequencing data showed that the dominant and rare OTUs ranged from 11 (Lake Aydat) to 26 (Lake Godivelle) and 281 (Lake Aydat) to 405 (Lake Godivelle), respectively (Fig. S2). Considering T-RFLP data, dominant T-RFs ranged from 7.8 (Lake Pavin) to 13 (Lake Bourget) and rare T-RFs ranged from 71 to 162 in the same lakes. These data suggest that rare taxa account for most of the small protists diversity.

Protists alpha-diversity

The mean number of total T-RFs varied in the euphotic zone from 83 in Lake Godivelle (oligotrophic) to 175 in Lake Bourget (mesotrophic). Lakes Pavin, Sep, Aydat, and Villerest revealed mean of total T-RFs of 134, 136, 163, and 155, respectively. By pyrosequencing, the results were different because the highest richness was detected for the most eutrophic system (Godivelle = 28.94, Pavin = 21.78, Sep = 25.44, Bourget = 27.72, Aydat = 18.9, and Villerest = 32.19) with a mean range of fluctuation of 17.2%. However, a similar pattern was obtained when considering only the richness index computed from pyrosequencing data restricted to dominant OTUs (Godivelle = 3.05, Pavin = 2.41, Sep = 2.16, Bourget = 3.12, Aydat = 1.59, and Villerest = 2.52).

To test the prediction of the island biogeography theory, richness data (estimated by two molecular methods) have been analyzed according to the area of the six lakes (Table 2). The inter-lake variations of the total, dominant and rare T-RFs were explained significantly ($P < 0.001$) by area. The area was also related to OTUs determined from pyrosequencing data but only for the dominants (Table 2). To test whether these relationships were not due to environmental factors, we processed a partial correlation allowing to study this relation when environmental factors do not vary. The environmental parameters (Table S1) were then represented by the first axis of a principal component analysis (PCA), synthesizing their variation with the richness. The first PCA axis associated with the T-RFLP experiments (24 samples) represented 38.0% of total variance and 46.1% for pyrosequencing data (six samples). The partial correlations were therefore also significant supporting that the significant effect measured can be attributed mainly to the main effect tested area. When the taxa–area relationships were significant, the slope 'z', a measure of the rate species turnover across space, varied between 0.06 and 0.10 for fingerprinting method and was equal to 0.14 for high-throughput sequencing (Table 2).

Protists beta-diversity

In terms of beta-diversity, the Table 3 shows the protists community composition (PCC) similarity between each lake by taking into account dominant and rare populations obtained from T-RFLP and pyrosequencing. Apparently, there is not congruence between both methods because, for example, lakes Aydat and Pavin shared

Table 2. Relation between small protists richness (from T-RFLP (T-RFs) and pyrosequencing (OTUs) data) and lake areas (calculated from the six French lakes). Data are log-transformed; freedom degrees associated with T-RF (six lakes × four dates) and OTUs are 22 and 4, respectively. Probability was computed from a one-sided test

Richness	Slope (z)	rs*	*P*	Partial rs†	*P*
T-RFs					
Total	0.09	0.79	< 0.001	0.79	< 0.001
Dominant	0.06	0.26	< 0.05	0.33	< 0.05
Rare	0.10	0.79	< 0.001	0.80	< 0.001
Pyrosequencing					
Total	0.03	0.08	NS	0.08	NS
Dominant	0.14	0.83	< 0.05	0.83	< 0.01
Rare	0.04	0.37	NS	0.38	NS

*rs: Spearman correlation.

†Partial Spearman correlation was computed with environmental parameters (first axis of a PCA) as constant.

Table 3. (a) % of common dominant (left panel) and rare (right panel) T-RFs for each pair of lakes (b) % of common dominant (left panel) and rare (right panel) OTUs for each pair of lakes

Lakes	Aydat	Bourget	Godivelle	Pavin	Sep	Villerest	
(a) T-RFs							
Aydat	100	36.52	25.00	53.73	58.85	55.3	
Bourget	6.67	100	15.85	29.67	31.09	34.04	
Godivelle	18.5	8.57	100	28.22	28.93	26.74	Rare
Pavin	21.74	21.43	17.24	100	51.38	54.55	
Sep	13.64	10.71	10.71	17.39	100	54.87	
Villerest	16.67	3.03	17.24	11.11	12.5	100	
			Dominant				
(b) OTUs (95%) from pyrosequencing							
Aydat	100	6.82	16.27	14.9	16.4	10.18	
Bourget	0.0	100	7.0	6.81	8.09	7.04	
Godivelle	4.55	0.0	100	13.74	14.6	13.2	Rare
Pavin	3.3	0.0	4.55	100	16.16	12.34	
Sep	0.0	0.0	0.0	2.7	100	19.47	
Villerest	0.0	2.38	0.0	0.0	11.11	100	
			Dominant				

21.7% of dominant T-RFs, whereas a value of 3.3% was obtained by high-throughput sequencing. Similarly, the common rare OTUs between each pairs of lakes were lower than the common rare T-RFs. However, a same pattern can be found with both methods: lakes shared more rare populations than dominants. The beta-diversity studied from the UniFrac distance (fraction of the total branch length in the phylogeny that is unique to any particular environments) shows that these distances between lakes obtained with pyrosequencing were significantly different ($P < 0.001$; Table S2).

The relative importance of local environmental factors and spatial distance on the protists repartition in the six French lakes has been analyzed with a Mantel test based on T-RFLP (time series) and pyrosequencing (single sampling) data (Table 4). The analysis showed a highly significant effect of geographic distance (spatial distribution at the regional scale) between sites for both rare and dominant T-RFs/OTUs, while the examined environmental variables do not seem to be significantly involved in the protists distribution.

Table 4. Effects of distance and environmental variables on protist community composition at regional scale

	Mantel		Partial mantel*	
Scale	*r*	*P*	*r*	*P*
T-RFs				
Total				
Spatial distribution	0.53	**< 0.01**		
Environmental variables	−0.15	0.87	−0.17	0.91
Dominant				
Spatial distribution	0.42	**< 0.001**		
Environmental variables	0.10	0.13	0.12	0.10
Rare				
Spatial distribution	0.49	**< 0.01**		
Environmental variables	−0.11	0.76	−0.12	0.81
OTUs from pyrosequencing				
Total				
Spatial distribution	0.83	**0.05**		
Environmental variables	−0.24	0.73	0.05	0.42
Dominant				
Spatial distribution	0.35	**0.03**	0.35	0.13
Environmental variables	0.19	0.23		
Rare				
Spatial distribution	0.84	**0.03**	0.18	0.33
Environmental variables	−0.18	0.66		

*The partial Mantel test holds spatial distribution constant.
Values in bold are significant at $P < 0.005$ (probability based on 999 permutations).
The environmental variables used at regional scale are reported in Table S1.

Discussion

Here, we analyzed, for the first time to our knowledge, alpha- and beta-diversity of the protist distribution patterns in view of geographic distances, lake areas, and habitat variables using two molecular methods including high-throughput sequencing.

Artifacts of taxonomic lumping, undersampling and unequal sampling could result in the incorrect conclusion about the spatial scaling of microbial biodiversity (Martiny *et al.*, 2006). Therefore, in our study, temperate lakes characterized by stable summer stratification, sampled in the epilimnetic zone, were selected across a regional scale. Moreover, our results are based on an equal

sampling effort and/or a time series, which allowed to take into account temporal variations in the community to analyze both spatial distribution and environment effects. Nolte *et al.* (2010) showed that seasonal abundance patterns of protists closely match their biogeographic distribution; temporal sampling is therefore basic for adequate diversity and species richness estimates. Finally, to avoid biases as much as possible, all lakes were sampled in the same way, and two molecular methods have been used for studying the spatial pattern of these microorganisms because the molecular methods used (fingerprinting, sequencing, etc.) can give different view on spatial patterns (Cho & Tiedje, 2000). In this study, the differences between T-RFLP results and OTUs determination can be due in part to the fact that a T-RF can correspond to diverse phylogenetic levels. It is clear that most of studies on protists diversity have undersampled their diversity greatly to this point. The application of novel high-throughput sequencing technologies, such as 454 pyrosequencing, has revealed a great diversity within bacterial and archaeal communities and allowed the access to the rarest OTUs (e.g. Sogin *et al.*, 2006; Galand *et al.*, 2009). However, these methods have been used for microbial eukaryotes in a really few studies so far in lacustrine ecosystems (e.g. Monchy *et al.*, 2011). High-throughput pyrosequencing technology address indeed methodological shortcomings by recovering uncommon and rare species, but the short read lengths of 454 sequences made it necessary to rely on the existing long rRNA gene sequences to establish taxonomic identities (Edgcomb *et al.*, 2011). Also, concerns remain about the role that sequencing errors may play in producing a distorted picture of the true complexity/richness of microbial communities (Kunin *et al.*, 2010). Analyses of pyrosequenced SSU rDNA fragments amplified from an artificial bacterial community suggested that sequencing errors result in richness estimates that are at least one order of magnitude too high (Quince *et al.*, 2009). However, this problem has been largely alleviated using computational tools to distinguish and filter out erroneous sequences (Quince *et al.*, 2009).

Spatial patterns of small protists were not linked to environmental factors

According to the cosmopolitan view of the microbial world, one might expect to find similar microbial community structure (richness, diversity, and composition) in similar habitats and differentiated microbial communities along an environmental gradient (Green & Bohannan, 2006). The highest richness found in the mesotrophic lake (lake Bourget) among the dominant populations could be therefore explained by its large area as well as by its intermediate trophic status, although most of the studies regarding protists (e.g. nanoflagellates) did not show a strong evidence of a link between diversity and trophic status (Arndt *et al.*, 2000; Auer & Arndt, 2001). However, phytoplankton and bacterial studies have reported that oligotrophic and eutrophic lakes presented lower diversity than mesotrophic lakes (Dodson *et al.*, 2000; Horner-Devine *et al.*, 2003). Experimental results also reported that phytoplankton diversity followed a hump-shape progression along a gradient of eutrophication. This diversity variation would allow a compromise between competition, predation, and accessibility of resources as well as other many ecological processes (e.g. Leibold, 1999). In this study, the use of partial correlations and the Mantel test allow to estimate the impact of geographic distance vs. environmental conditions on assemblage composition (Martiny *et al.*, 2006), factors rarely taken into account simultaneously. The results clearly showed that variations of the alpha- and beta-diversity were significantly influenced by the geographic distance between lakes or areas rather than the environmental factors analyzed, such as bottom-up factors (i.e. nutrients) or potential predators (i.e. HNF and metazooplankton). However, even though the parameters explored are those commonly considered when attempting to explain the spatial partitioning of aquatic microorganisms, we did not analyze all potential controlling factors. For example, Schiaffino *et al.* (2011) showed that light penetration and DOC had a structuring effect on microorganism populations in 45 lakes. Overall, our results contradict the hypothesis of a general microbial cosmopolitanism. These patterns have already been observed for bacterial communities (Reche *et al.*, 2005; Martiny *et al.*, 2011). In addition, this analysis performed with all T-RFLP data (monthly sampling) showed therefore that temporal variations in the composition of the small protists community do not affect the importance of geographic distances. We can therefore hypothesize that the use of a single sampling per lake (pyrosequencing) could be enough to analyze the spatial repartition of lacustrine protists at different spatial scales.

Distribution of dominant small protists was linked to lake area and distance between lakes

MacArthur and Wilson's theory of island biogeography is among the most well-known process–based explanations for the distribution of species richness (alpha-diversity) and has been applied for microorganisms' biogeography studies (e.g. Reche *et al.*, 2005; Logue *et al.*, 2012). It helps understanding the taxa–area relationship, a fundamental pattern in ecology and an essential tool for conservation. Most of what we know of taxa–area curves is derived from analyses of terrestrial systems, but lakes and

ponds can be considered as discrete habitats with definable borders that are comparable in some ways to oceanic islands (Dodson, 1992). This theory has been indeed already tested for zooplankton, phytoplankton, and bacteria communities in lakes (Dodson, 1992; Smith *et al.*, 2005; Reche *et al.*, 2005; Logue *et al.*, 2012). In our study, the linear relationships observed between lake area and protists richness (determined by the fingerprinting method) were significant for total, dominant, and rare T-RFs, whereas the same relationships determined by pyrosequencing involved mainly the dominant species (i.e. OTUs). The TAR determined in this study varied with the molecular method used, as already highlighted (e.g. Zhou *et al.*, 2008), and therefore with the taxonomic resolution. However, all methods suggested that the spatial distribution of dominant protists follows the TAR of the island biogeography theory.

Significant TAR was recently found for the richness of phytoplankton (Smith *et al.*, 2005) as well as bacteria (Reche *et al.*, 2005). However, as emphasized above, this theory seems restricted to the dominant taxa (i.e. OTUs) and to a specific range of lake area. The slopes of TAR varied also with the method used as already showed (Zhou *et al.*, 2008). To compare to bibliographic data, we focused on the 'z' values determined at the 'OTUs' level. The z value determined with pyrosequencing (0.14) is quite far from the TAR determined for ectomycorrhizal fungi (Peay *et al.*, 2007) but is close to those found for other planktonic organisms such as zooplankton [$z = 0.094$ (Dodson, 1992)] and phytoplankton ($z = 0.114$, Smith *et al.*, 2005). For microorganisms, the lowest 'z' values were recorded for benthic ciliates ($z = 0.043$, Finlay *et al.*, 1998) or bacteria ($z = 0.040$, Horner-Devine *et al.*, 2004). Finally, our results contradict for dominant populations the advocates of microorganism cosmopolitan distribution, which suggests that microorganisms should be characterized by a flat TAR. Moreover, if the richness (alpha-diversity) seems to be constant for total OTUs (dominants and rares), PCC shows important changes. Therefore, even if the pyrosequencing data highlighted that the alpha-diversity did not vary with the habitat size, beta-diversity was strongly associated with distance for total, dominant, and rare OTUs. Hillebrand *et al.* (2001) showed also a distance decay relationship for diatoms and ciliates, but these authors did not consider the putative effects of environmental parameters.

Biogeographic patterns of the rarest protists taxa

The other component of the community, named 'the rare biosphere', comprises a very high number of rare species that contains most of the diversity (Sogin *et al.*, 2006). Our results show that, while the TAR cannot be applied to rare OTUs (pyrosequencing data), the composition (beta-diversity) was linked to the geographic distances for the rarest T-RFs and OTUs. Thus, rare community composition varies between ecosystems but not the richness. Similarly, Galand *et al.* (2009) reported that the rare biosphere of the archaeal community followed patterns similar to those of the most abundant members of the community and has a biogeographic pattern. Based on the assumption that rare OTUs are taxa with low abundance, we assume that rare taxa could be more impacted by dispersal limitation because the probability of immigrating and growing in a new ecosystem is limited compared to abundant taxa (Green *et al.*, 2004; Weisse, 2008). The 'rare biosphere' has traditionally been thought to indicate the presence of a seed bank of potential new colonizers, according to the 'everything is everywhere' hypothesis. However, the Mantel test showed no correlation between environmental variables and the rare T-RFs/OTUs. Galand *et al.* (2009) showed that regardless of spatial or temporal scales, most of the rare phylotypes are always rare within an ecosystem and the few rare phylotypes that are sometimes detected as abundant represent traces of phylotypes that are highly abundant in some habitats.

At a larger geographic scale: what do we learn from public database?

To extend our discussion to a more global scale (across continent), we used sequences available on the same planktonic size fraction from public database. These sequences were obtained by the traditional cloning–sequencing method (11 worldwide lakes). Although with this approach the sampling is far from exhaustive and many more taxa (especially rare) might be present at the

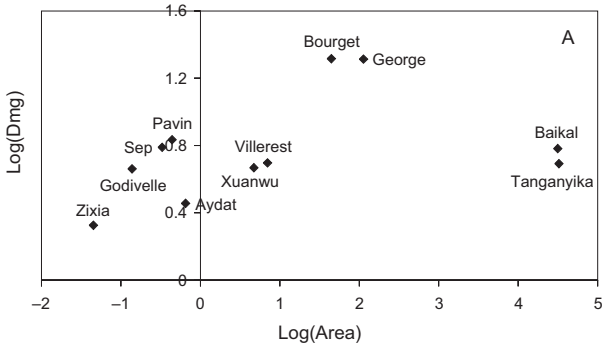


Fig. 1. Relationship between lake areas and richness expressed by Margalef index (Dmg) calculated from cloning–sequencing method (six French lakes + five worldwide lakes = Baikal (JN547261-JN547327), George (AY919677-AY919829), Tanganyika (GU290066-GU290116), Zixia (FJ939033-FJ939124), Xuanwu (FJ939033-FJ939124)).

sampling sites, statistical analysis (UniFrac analysis) showed 18S rRNA gene OTUs compositions significantly different among most of lakes, and this difference does not seem to be related to the trophic status at any spatial scales (data not shown). Such results were expected because the effect of dispersal limitation would occur at the largest geographic scales (across continents) rather than at regional scales where environment selection could theoretically be predominant (Martiny *et al.*, 2006).

At the global scale, a significant linear relationship with lake areas ($z = 0.24$, $P < 0.001$, $n = 9$) is obtained but only when using the smaller lake areas (lower than 114 km² corresponding the lakes Bourget and George; Fig. 1). The Fig. 1 shows a type IV curve typically calculated as a linear regression on a log–log scale from samples of island-like habitats (e.g. lakes) with possible decreasing properties (Scheiner, 2003). In a recent study conducted on Bacteria richness in 14 Swedish lakes, Logue *et al.* (2012) showed by analyzing 454 pyrosequencing data that taxa–area relationships were negative. These data suggest that the island model is therefore nonrestricted to a positive TAR, and the area range of ecosystems chosen must be large enough for assessing the model type.

Conclusion

The use of several approaches (including 454-pyrosequencing) allows to reduce bias due to inadequate sampling of rare taxa (Woodcock *et al.*, 2006) and the difficulty of delineating microbial species (Horner-Devine *et al.*, 2004). Our study clearly shows distance–decay pattern in terms of beta-diversity for rare and dominant small protists. The spatial distribution of dominant taxa followed both predictions of the island biogeography theory following the hypothesis that lakes act as ecological islands, not only for macroorganisms but also for microorganisms (Rengefors *et al.*, 2012). Unlike to our hypothesis, geographic (i.e. distances) effect seems to be the only one shaping the PCC at global and regional scales suggesting that one cause of the PCC differentiation is due to physical barriers. However, population differentiation may also be due to biological barriers. It is, however, difficult to assess speciation and extinction rates of microorganisms *in situ* especially when considering heterogeneous group of organisms, as protists, which comprise a high diversity of functional roles. There is likely no general rule for the biogeography of microorganisms, and results seem to change according to the group studied. A scenario proposed by Rengefors *et al.* (2012) to explain dinoflagellates differentiation in lakes is the 'Monopolization Hypothesis', which states that genetic differentiation can be explained by rapid population growth after historical founder events, enhanced by the presence of a large resting propagule bank providing a powerful buffer against newly invading genotypes (De Meester *et al.*, 2002). Another hypothesis is that the richness/diversity of the smallest protist studied here is not truly independent of the diversity and structure of other (larger) planktonic compartments due to competition or parasitism. In this case, the high importance of parasitic groups recently highlighted among these protists in lacustrine ecosystems (e.g. Lepère *et al.*, 2008) might explain an evolutionary drift due to 1) a host–parasite co-evolution in a lake, 2) the absence or extinction of the host and parasites have to change their primary hosts (common process in parasitic interactions) and such events often lead to speciation (Zietaria & Lumme, 2002). For example, the highly differentiated population of freshwater diatoms described by Evans *et al.* (2009) could be also explained by a host–parasite relation with chytrids (Jobard *et al.*, 2010; Rasconi *et al.*, 2011) or Cryptomycota (Jones *et al.*, 2011). Therefore, we need to extend our knowledge of eukaryote diversity and species association in lakes (e. g. host–parasite relationship, syntrophy, etc.) to highlight a biogeography of co-occurrence.

Acknowledgements

This study was supported by financial aids from INSU EC2CO. We would like to thank Pr. N. Melnik and Dr. O. Belykh for the hosting and cruise organization during the lake Baikal sampling.

References

Arndt H, Dietrich D, Auer B, Cleven EJ, Grafenham T, Weitere M & Mylnikov AP (2000) Functional diversity of heterotrophic flagellates in aquatic ecosystems, Systematics association Special volume no. 59. *Flagellates: Unity, Diversity and Evolution* (Leadbeater BSC & Green JC, eds), pp. 240–268. Taylor and Francis, London.

Auer B & Arndt H (2001) Taxonomic composition and biomass of heterotrophic flagellates in relation to lake trophy and season. *Freshw Biol* **46**: 959–972.

Baas-Becking LGM (1934) *Geobiologie of Inleiding Tot de Milieukunde*. W.P. Van Stockum & Zoon, The Hague, the Netherlands (in Dutch)

Caron D (2009) Past President's Address: Protistan Biogeography: Why All The Fuss? *J Eukaryot Microbiol* **56**: 105–112.

Cho JC & Tiedje JM (2000) Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. *Appl Environ Microbiol* **26**: 5448–5456.

De Meester L, Gomez A, Okamura B & Schwenk K (2002) The Monopolization Hypothesis and the dispersal-gene flow paradox in aquatic organisms. *Acta Oecol* **23**: 121–135.

Dodson S (1992) Predicting crustacean zooplankton species richness. *Limnol Oceanogr* **37**: 848–856.

Dodson SI, Arnott SE & Cottingham KL (2000) The relationship in lake communities between primary productivity and species richness. *Ecology* **81**: 2662–2679.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Edgcomb V, Orsi W, Bunge J *et al.* (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J* **5**: 1344–1356.

Evans KM, Chepurnov VA, Sluiman HJ *et al.* (2009) Highly differentiated populations of the freshwater diatom *Sellaphora capitata* suggest limited dispersal and opportunities for allopatric speciation. *Protist* **160**: 386–396.

Finlay BL & Fenchel T (2004) Cosmopolitan metapopulations of free-living microbial eukaryotes. *Protist* **155**: 237–244.

Finlay BJ, Esteban GF & Fenchel T (1998) Protozoan diversity: converging estimates of the global number of free-living ciliate species. *Protist* **149**: 29–37.

Finlay BJ, Esteban GF, Brown S, Fenchel T & Hoef-Emden K (2006) Multiple cosmopolitan ecotypes within a microbial eukaryote morphospecies. *Protist* **157**: 377–390.

Foissner W (2006) Biogeography and dispersal of micro-organisms: a review emphasizing protists. *Acta Protozool* **45**: 111–136.

Galand PE, Casamayor EO, Kirchman DL & Lovejoy C (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. *P Natl Acad Sci USA* **106**: 22427–22432.

Good IL (1953) The population frequencies of species and the estimation of population parameters. *Biometrika* **40**: 237–264.

Green J & Bohannan BJM (2006) Spatial scaling of microbial biodiversity. *Trends Ecol Evol* **21**: 501–507.

Green JL, Holmes AJ, Westoby M, Oliver I, Briscoe D, Dangerfield M, Gillings M & Beattie AJ (2004) Spatial scaling of microbial eukaryote diversity. *Nature* **432**: 747–750.

Hill TCJ, Walsh KA, Harris JA & Mofett BF (2003) Using ecological diversity measures with bacterial communities. *FEMS Micobiol Ecol* **43**: 1–11.

Hillebrand H, Watermann F, Karez R & Berninger UG (2001) Differences in species richness patterns between unicellular and multicellular organisms. *Oecologia* **126**: 114–124.

Horner-Devine MC, Leibold MA, Smith VH & Bohannan BJM (2003) Bacterial diversity patterns along a gradient of primary productivity. *Ecol Lett* **6**: 613–622.

Horner-Devine MC, Lage M, Hughes JB & Bohannan BJM (2004) A taxa-area relationship for bacteria. *Nature* **432**: 750–753.

Jobard M, Rasconi S & Sime-Ngando T (2010) Diversity and functions of microscopic fungi: a missing component in pelagic food webs. *Aquat Sci* **72**: 255–268.

Jones MD, Forn I, Gadelha C, Egan MJ, Bass D, Massana R & Richards TA (2011) Discovery of novel intermediate forms redefines the fungal tree of life. *Nature* **11**: 474–479.

Kunin V, Engelbrektson A, Ochman H & Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118–123.

Lefranc M, Thénot A, Lepère C & Debroas D (2005) Genetic diversity of small eukaryotes in lakes differing by their trophic status. *Appl Environ Microbiol* **71**: 5935–5942.

Leibold MA (1999) Biodiversity and nutrient enrichment in pond plankton communities. *Evol Ecol Res* **1**: 73–95.

Lepère C, Boucher D, Jardillier L, Domaizon I & Debroas D (2006) Succession and regulation factors of small eukaryote community composition in a lacustrine ecosystem (Lake Pavin). *Appl Environ Microbiol* **72**: 2971–2981.

Lepère C, Domaizon I & Debroas D (2007) Community composition of lacustrine small eukaryotes in hyper-eutrophic conditions in relation to top-down and bottom-up factors. *FEMS Microbiol Ecol* **61**: 483–495.

Lepère C, Domaizon I & Debroas D (2008) Unexpected importance of potential parasites in the composition of the freshwater small-eukaryote community. *Appl Environ Microbiol* **74**: 2940–2949.

Lindstrom ES & Langenheder S (2012) Local and regional factors influencing bacterial community assembly. *Environ Microbiol Rep* **4**: 1–9.

Logue JB, Langenheder S, Andersson AF, Bertilsson S, Drakare S, Lanzén A & Lindström ES (2012) Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species–area relationships. *ISME J* **6**: 1127–1136.

Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C & Moreira D (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**: 603–607.

Lozupone C & Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.

MacArthur RH & Wilson EO (1967) *The Theory of Island Biogeography*. Princeton University Press, Princeton, New Jersey, USA.

Mangot JF, Domaizon I, Taib N, Marouni N, Duffaud E, Bronner G & Debroas D (2013) Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. *Environ Microbiol*. doi: 10.1111/1462-2920.12065 [Epub ahead of print].

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res* **27**: 209–220.

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL *et al.* (2006) Microbial biogeography: Putting microorganisms on the map. *Nat Rev Microbiol* **4**: 102–112.

Martiny JBH, Eisenb JA, Pennc K, Allisona SD & Horner-Devine MC (2011) Drivers of bacterial β-diversity depend on spatial scale. *P Natl Acad Sci USA* **108**: 850–7854.

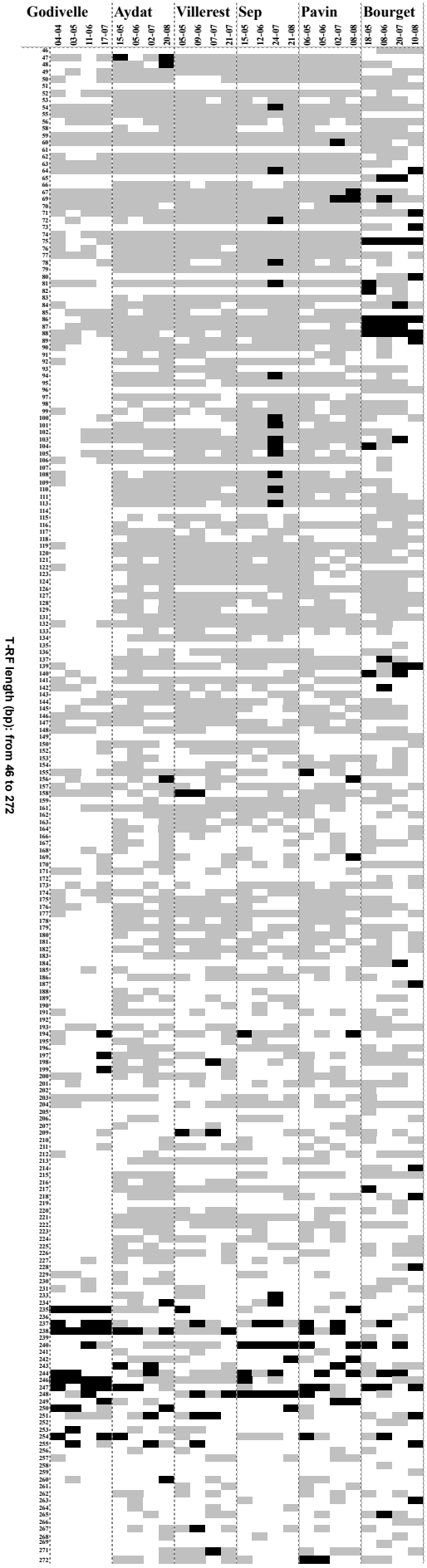
Monchy S, Sanciu G, Jobard M, Rasconi S, Gerphagnon M, Chabé M, Cian A, Meloni D, Niquil N & Christaki U (2011)

Exploring and quantifying fungal diversity in freshwater lake ecosystems using rDNA cloning/sequencing and SSU tag pyrosequencing. *Environ Microbiol* **13**: 1433–1453.
Nolte V, Pandey RV, Jost S, Medinger R, Ottenwalder B, Boenigk J & Schlötterer C (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol Ecol* **19**: 2908–2915.
Papke RT & Ward DM (2004) The importance of physical isolation to microbial diversification. *FEMS Microbiol Ecol* **48**: 293–303.
Peay KG, Bruns TD, Kennedy PG, Bergemann SE & Garbelotto M (2007) A strong species-area relationship for eukaryotic soil microbes: island size matters for ectomycorrhizal fungi. *Ecol Lett* **10**: 470–480.
Pedros-Alio C (2006) Marine microbial diversity: can it be determined? *Trends Microbiol* **14**: 257–263.
Pinel-Alloul B & Ghadouani A (2007) Spatial heterogeneity of planktonic microoraganisms in aquatic systems: mutiscale patterns and processes. *The Spatial Distribution of Microbes in the Environment* (Franklin RB & Mills AL, eds), pp. 203–310. Springer, Netherlands.
Pither J (2007) Comment on "Dispersal Limitations Matter for Microbial Morphospecies". *Science* **316**: 1124.
Price MN, Dehal PS & Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641–1650.
Quince C, Lanzen A, Curtis TP *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**: 639–641.
Rasconi S, Jobard M & Sime-Ngando T (2011) Parasitic fungi of phytoplankton: ecological roles and implications for microbial food webs. *Aquat Microb Ecol* **62**: 123–137.
Reche I, Pulido-Villena E, Morales-Baquero R & Casamayor EO (2005) Does ecosystem size determine aquatic bacterial richness? *Ecology* **86**: 1715–1722.
Rengefors K, Logares R & Laybourn-Parry J (2012) Polar lakes may act as ecological islands to aquatic protists. *Mol Ecol* **21**: 3200–3209.
Richards TA, Vepritskiv AA, Gouliamova DE & Nierzwicki-Bauer SA (2005) The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. *Environ Microbiol* **7**: 1413–1425.
Scheiner SM (2003) Six types of species-area curves. *Glob Ecol Biogeogr* **12**: 441–447.
Schiaffino MR, Unrein F, Gasol JM, Massana R, Balagué V & Izaguirre I (2011) Bacterial community structure in a latitudinal gradient of lakes: the roles of spatial versus environmental factors. *Freshw Biol*. doi: 10.1111/j.1365-2427.2011.02628.x
Smith VH, Foster BL, Grover JP, Holt RD, Leibold MA & DeNoyelles F (2005) Phytoplankton species richness scales consistently from laboratory microcosms to the world's oceans. *P Natl Acad Sci USA* **102**: 4393–4396.
Sogin ML, Morisson HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *P Natl Acad Sci USA* **103**: 12115–12120.
Taylor JW, Turner E, Townsend JP, Dettman JR & Jacobson D (2006) Eukaryotic microbes, species recognition and the geographic limits of species: examples from the kingdom Fungi. *Philos Trans R Soc Lond B Biol Sci* **361**: 1947–1963.
Telford RJ, Vandik V & Birks HJB (2006) Dispersal limitations matter for microbial morphospecies. *Science* **312**: 1015.
Weisse T (2008) Distribution and diversity of aquatic protists: an evolutionary and ecological perspective. *Biodivers Conserv* **17**: 243–259.
Woodcock S, Curtis TP, Head IM, Lunn M & Sloan WT (2006) Taxa-area relationships for microbes: the unsampled and the unseen. *Ecol Lett* **9**: 805–812.
Wuyts J, Perriere G & Van De Peer Y (2004) The European ribosomal RNA database. *Nucleic Acids Res* **32**: D101–D103.
Yang J, Smith HG, Sherratt TN & Wilkinson DM (2010) Is there a size limit for cosmopolitan distribution in free-living microorganisms? A biogeographical analysis of testate Amoebae from polar areas. *Microb Ecol* **59**: 635–645.
Zhou J, Kang S, Schadt CW & Garten CT (2008) Spatial scaling of functional gene diversity across various microbial taxa. *P Natl Acad Sci USA* **105**: 7768–7773.
Zietaria MS & Lumme J (2002) Speciation by host switch and adaptive radiation in a fish parasite genus Gyrodactylus (Monogenea, Gyrodactylidae). *Evolution* **56**: 2245–2458.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Spatio-temporal dynamics of rare (in grey) and dominant (in black) T-RFs.
Fig. S2. Spatial distribution of dominant OTUs (in black) determined by pyrosequencing.
Table S1. Environmental parameters measured in the six French lakes.
Table S2. Unifrac distances for comparing PCC in each lake to each other lake by pyrosequencing.
Data S1. Pyrosequencing data analysis.



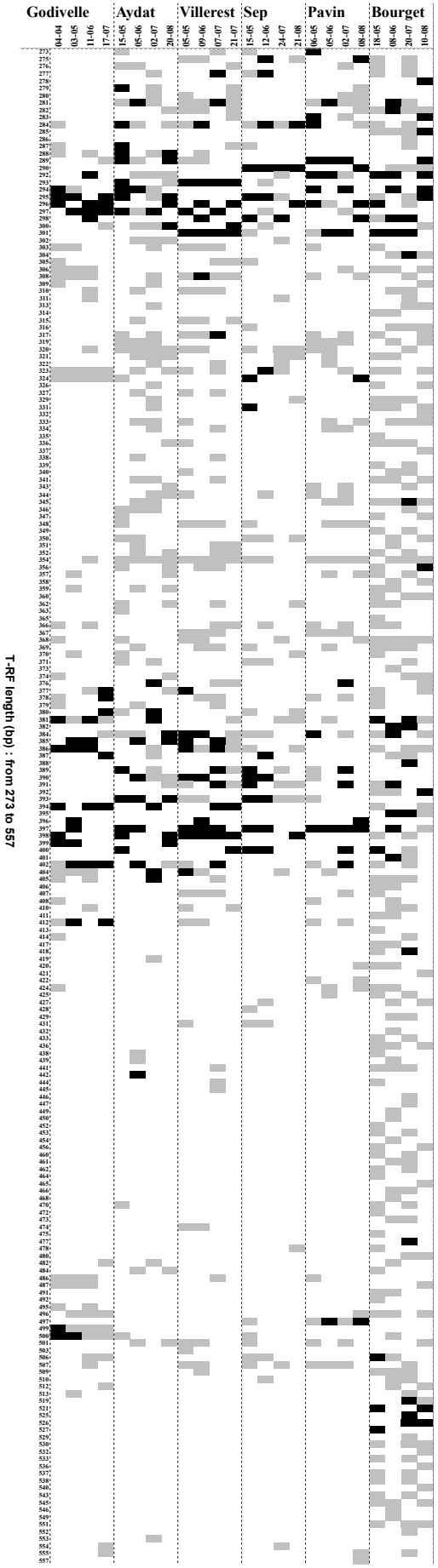


Fig S1: Spatio-temporal dynamics of rare (in grey) and dominant (in black) T-RFs

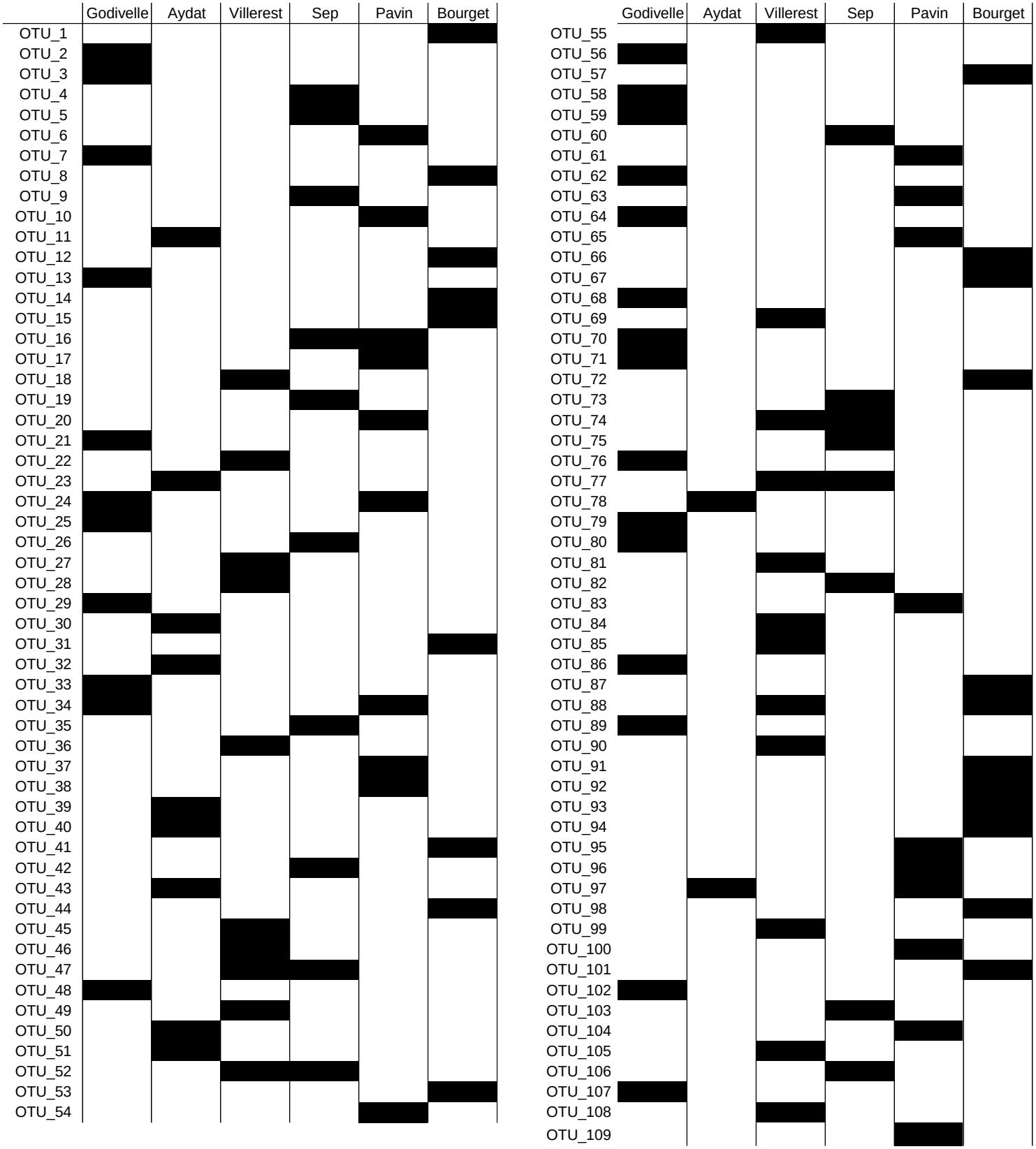


Fig. S2: Spatial dynamics of dominant OTUs (in black) determined by pyrosequencing

Lakes	$N\text{-}NH_4$ ($mgN.L^{-1}$)	$N\text{-}NO_3$ ($mgN.L^{-1}$)	$P\text{-}PO_4$ ($mgP.L^{-1}$)	Temp (C°)	Chl *a* ($\mu g.L^{-1}$)	Secchi (m)	Cladocerans (ind. L^{-1})	Copepods (ind. L^{-1})	Rotifers (ind. L^{-1})	HNF (10^3 Cell. mL^{-1})	Bacteria (10^6 Cell. mL^{-1})
Godivelle	0.01	0.32	0	13.2	0.032	7.65	0.25	0.48	0.16	0.09	1.20
Pavin	0.002	0.03	0.028	13.0	1.5	5.50	1.0	4.8	34.2	0.15	4.74
Sep	0.05	0.55	0.034	18.5	10.3	3.13	10.5	103.3	96.0	25.29	10.41
Bourget	0.005	0.25	0.003	19.5	0.5	7.97	3.35	10.35	8.70	1.64	2.72
Aydat	0.01	0.029	0.012	21.2	11.23	1.85	19.5	16	154.5	5.08	6.82
Villerest	0.08	1.16	0.039	19.0	12.14	2.89	12.1	40.0	44.7	2.05	6.71

Table S1: Environmental parameters measured in the 6 French lakes. Values are means for the period studied (see M&M) HNF: heterotrophic nanoflagellates; ind: individuals; Chl *a:* chlorophyll *a;* Secchi: water clarity

Supplementary text S1: Pyrosequencing data analysis

Reference database

Experimental sequences analyzed through pyrosenquencing were compared with a dedicated database of reference sequences extracted from the SSURef 108 database from the SILVA database project, which offers taxonomic information, quality assessment and a curated alignment of SSU rRNA sequences (Pruesse *et al*. 2007). The database only includes sequences with more than 1200 bp, quality score > 75%, and pintail value > 50 according to the SILVA classification. In this implementation, sequences that belong to the *Eukarya* domain were kept, corresponding to 15,419 sequences from microeukaryotes. Sequences from pluricellular organisms (*Metazoa*, *Rhodophyta* and *Streptophyta*), plus one bacterial sequence and one archaeal sequence were added to allow the detection of experimental sequences affiliated to them.

To speed up the phylogenetic processing, the reference alignments were split into 24 phyletic groups, and groups with more than 1000 sequences were clustered using complete linkage clustering at 97%. For each phyletic group, an outgroup containing one sequence from each of the other phyletic groups plus two metazoan sequences were added to the alignment to root the phyletic tree to be produced, and to specify the relatedness of early diverging sequences from the root of the group. The sequences from each phyletic group together with the outgroup sequences were retrieved from the SILVA alignment using ARB, and then trimmed to remove vertical gaps. Besides these files, an HMM profile was built from each of them using HMMbuild from the HMMER package (Eddy, 1998). A taxonomy file containing EMBL taxonomy of each sequence of the reference database was also generated.

Processed sequences

1- By using PANGEA (Giongo *et al*. 2010) functionalities, short sequences (<200 bp) and sequences with low quality (≤ 23) were removed.

2- Clean reads were then clustered with UCLUST (Edgar, 2010) with a clustering threshold of 95% which allows an approximate species-level distinctions (Caron *et al*. 2009). Moreover, our *in silico* analysis (results no shown) demonstrated that an identity threshold of 95% applied on the V4-V5 hypervariable regions of the 18S rRNA allows a similar taxonomic affiliation as a 97% cut-off that is applied on the entire small subunit rRNA gene and generally used for clustering pyrosequencing reads.

3- The OTUs were compared against the reference database with USEARCH (Edgar, 2010). Then,

following the taxonomy of its best hit, each sequence was appended to a phyletic group, together with its five best hits. The query sequences are sorted according to their assignment.

4- Homologous reads have been then assigned to phyletic groups, they were aligned with the referenced sequences of the corresponding profile using HMMalign (Eddy, 1998). Next, a phylogenetic tree was build for each phyletic profile, using FASTTREE (Price *et al*. 2009), with the Jukes-Cantor + Cat model and a bootstrap threshold of 100. If the number of reads belonging to one phyletic group exceeds 30,000, then this group was split into files containing fewer than 30,000 reads.

5- Finally, the trees were parsed to generate files containing the taxonomy of the inserted sequences. Taxonomy assessment was inferred by nearest neighbor. For each query sequence, all the nodes containing it were scanned from the most recent to the deepest. The closest neighbor was defined as the first referenced sequence starting from the lowest node. The query sequence will acquire the complete taxonomy of its closest neighbor.

6- The built trees were used to compute Unifrac distance.

The package used for this analysis (named PANAM) can be obtained from http://code.google.com/p/panam-phylogenetic-annotation/. It comprises the reference sequences database, the taxonomy file and reference profile alignments (Mangot *et al*. 2012).

Caron, D.A., Countway, P.D., Savai, P., Gast, R.J., Schnetzer, A., Moorthi, S.D. *et al.* (2009). Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl. Environ. Microbiol.*, 75, 5797-5808.

Eddy SR: Profile hidden Markov models. (1998). *Bioinformatics* 14, 755- 63.

Edgar RC: Search and clustering orders of magnitude faster than BLAST. (2010). *Bioinformatics* 26, 2460-2461.

Giongo, A., Crabb, D.B., Davis-Richardson, A.G., Chauliac, D., Mobberley, J.M., Gano, K.A., Mukherjee, N., Casella, G., Roesch, L.F., Walts, B. *et al*. (2010). PANGEA: pipeline for analysis of next generation amplicons. *ISME J.*, 4, 852-861.

Mangot, J.F., Domaizon., I, Taib, N., Marouni, N., Duffaud, E., Bronner, G., Debroas, D. (2013). Short term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. *Environ. Microb*., doi: 10.1111/1462-2920.12065.

Price, M.N., Dehal, P.S., & Arkin, P.A. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.*, 26, 1641-1650.

Pruesse, E., Quast, C., Knittel., K, Fuchs, B., Ludwig, W., Peplies, J., & Glockner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence

data compatible with ARB. *Nucl. Acids. Res.*, 35, 7188-7196.